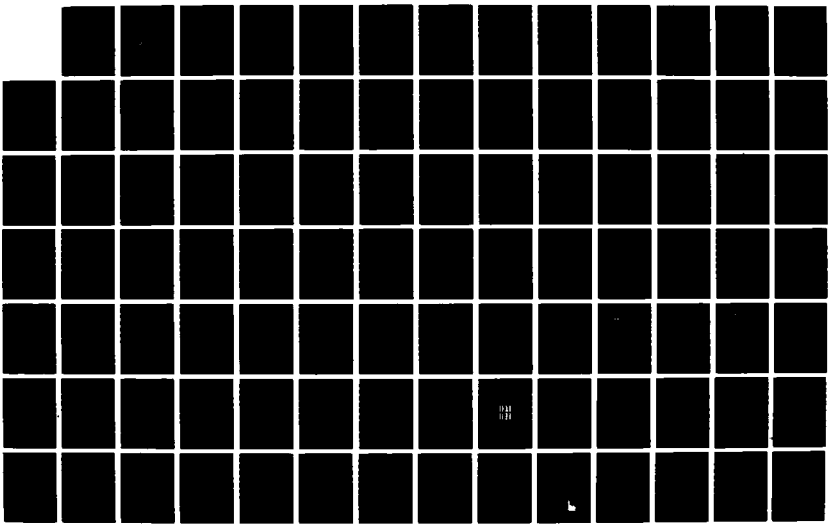
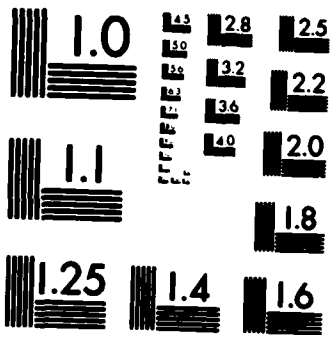
AD-A188 009 PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP HELD IN COLLEGE..(U) TEXAS A AND M UNIV COLLEGE STATION R B LIVINGSTON AUG 85 1/5

UNCLASSIFIED DAMD-17-85-G-5042 F/G 6/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(1)

VOLUME III: BACKGROUND PAPERS SUBMITTED BY PARTICIPANTS

PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP

Held at
Texas A&M University
College Station, TX 77843
August 10-16, 1985

Cosponsored by the
United States Army Medical Research and Development Command,
Scripps Clinic and Research Foundation,
Texas A&M University,
University of California, San Diego, and
Washington University

Supported by Grant #DAMD 17-85-G-5042
with the University of California, San Diego,
Robert B. Livingston, M.D., Principal Investigator
Grantor: U.S. Army Medical Research & Development Command
Fort Detrick, Frederick, MD 21701-5012

SECURITY CLASSIFICATION OF THIS PAGE

AD-A182209

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Texas A&M University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) College Station, Texas 77843	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Medical Research & Development Command 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAMD17-85-G-5042
8c. ADDRESS (City, State, and ZIP Code) Fort Detrick, Frederick, MD 21701-5012	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Brain Mapping Machine Design Workshop

12. PERSONAL AUTHOR(S)
Robert B. Livingston

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final Vol III	FROM 8/1/85 TO 5/31/86	1985 August	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Mrs. Virginia Miller	22b TELEPHONE (Include Area Code) 301/663-7325 22c. OFFICE SYMBOL SGRD-RMI-S

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

Contract No. DAMD17-85-G-5042

Title: Brain Mapping Machine Design Workshop

Robert B. Livingston, M.D.

Texas A & M University
College Station, Texas 77843

August 1985

Final Report, Volume III

Prepared For: U.S. Army Medical Research and Development Command
Fort Detrick, Frederick, Maryland 21701-5012

Distribution Statement:

Approved for public release; distribution unlimited

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

VOLUME III: BACKGROUND PAPERS SUBMITTED BY PARTICIPANTS

PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP

Held at
Texas A&M University
College Station, TX 77843
August 10-16, 1985

Cosponsored by the
United States Army Medical Research and Development Command,
Scripps Clinic and Research Foundation,
Texas A&M University,
University of California, San Diego, and
Washington University

Supported by Grant #DAMD 17-85-G-5042
with the University of California, San Diego,
Robert B. Livingston, M.D., Principal Investigator
Grantor: U.S. Army Medical Research & Development Command
Fort Detrick, Frederick, MD 21701-5012

PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP

Held at Texas A&M University
College Station, Texas 77843
August 10-16, 1985

VOLUME III: BACKGROUND PAPERS SUBMITTED BY PARTICIPANTS

Partial Contents:

0. General

WORKSHOP REPORT ON

APPLICATIONS OF ADVANCED COMPUTERS IN LIFE SCIENCES

Held on December 10-12, 1984 at the
Airlie Foundation, Airlie, Virginia

Contributions from the participants edited by section chairpersons:

Stephen Harrison.....Harvard University
Wayne Hendrickson.....Columbia University
Dean Hillman.....New York University Med. Ctr.
Rodolfo Llinas.....New York Univ. Med. Ctr.
Charles Peskin.....New York University

NSF Associates:

James Larimer..Sensory Physiology and Perception
John Wooley.....Instrumentation Program

Conference Director:

Dr. Dean E. Hillman
Dept. of Physiology & Biophysics
New York University Medical Ctr.
550 First Ave, New York, NY 10016
(212) 340-5417

Supported by a grant from the National Science Foundation
to the New York University Medical Ctr. (R. Llinas, Project Director)

SUMMARY

A workshop, sponsored by NSF, was held on December 10-12, 1984, to discuss advanced computer applications in the life sciences. The key issues were:

1. What are present and potential applications of supercomputers in the life sciences?

2 What areas will benefit most from supercomputing?

3. What other advanced computing technologies are important for progress in the life sciences?

4. What are the barriers to effective use of supercomputers and other advanced computing technologies and how can they be overcome?

The group agreed on the following general conclusions:

1. There are important immediate applications in: a) studies of macromolecular structure and dynamics, b) mathematical modeling of dynamic physiological systems, c) image processing, d) population biology and e) data bases of several sorts. In many of these fields, we anticipate that major questions will be answered in the next five years through use of supercomputers.

2. Substantially increased use of supercomputers can be anticipated as initial problems are solved, communications improve, and sophistication grows through "hands-on" experience.

3. Effective use of supercomputers by life scientists depends critically on ready access to broad-band telecommunications and to outstanding computer graphics.

4. Effective use of supercomputers by life scientists and increased sophistication and training in computational approaches depend crucially on maintenance of excellent local computing capabilites (large mini computers and graphics) at the highest level.

5. Developments such as special-purpose computers and array processors offer exciting opportunities as complementary approaches in many applications.

SYNOPSIS OF WORKSHOP

General Conclusions

Supercomputers are applicable to a very wide range of research approaches in life sciences. Currently, a number of these disciplines have taken advantage of the computational speed and large array memories inherent in these machines. Many algorithms and codes for biological analysis are readily vectorizable to optimize supercomputer cycle utilization. Because of the lack of adequate communications available to life scientists, effective use of these facilities requires residence near the site or alternatively the investigator must operate in a remote, batch-mode environment with turn around time for results taking days. This limits the number of potential applications to about 1/4 -1/3 the number of studies that could effectively gain from supercomputer implementation. The requirement is for telecommunications having a range of 9600 to over 56K baud band width. Many applications require even higher band widths for tranferring data into supercomputer centers and then graphically displaying the output back to the laboratory.

Three general categories of analysis in life sciences would greatly benefit from computational power of supercomputers by lifting barriers to analysis.

Analysis and modeling of dynamic processes and related structures *requiring only batch mode processing should be immediately instituted at advanced computational installations.* One of the principal biological applications is modeling of dynamic processes using available biological parameters. The range of applications extends from determining molecular conformation to dynamics of organs and body functions. Examples of ionic, molecular and macromolecular dynamics are protein folding, enzyme action, antigen-antibody reactions, toxic substance binding, channel activity, cellular communications (i.e., in cytogenesis and intra- and interneuronal activity) and functional mechanisms within and between macromolecules and organelles. Applications to larger structures include cell and interstitial tissue growth as well as displacement and interaction of body fluids with macromolecules, organelles and organs.

Analysis of large data bases, *where a number of correlations or transformations must be obtained, is a critical area requiring both continued support of computational projects and new developments in the area of telecommunications and local graphic capabilities.* These data are from crystallography, microscopy (light and electron), physiology and data base compilations. Analysis of large data bases, such as data banks of macromolecular atomic coordinates and information from population genetics requires extensive searches of various parameters. These studies are immediately applicable to supercomputers provided that sufficient data storage is available at the site. Already crystallographic data are being analyzed extensively on Class VI computers. Physiological analysis, on the other hand, generates massive amounts of data, which arrive at high rates in real time. These data are needed for analysis of events occurring between ions, molecules, macromolecules, organelles and cells within a life sustaining environment. In areas of neuronal circuit analysis, both spatial and temporal analysis of multiple elements at high band widths are required for extended periods of time. Effective means of communications between laboratories and the supercomputer centers are essential in the majority of these applications.

Development of applications algorithms and software *for supercomputers in all areas of life sciences must be stimulated.* The testing of new algorithms and codes for applications that require long runs or large data arrays is possible by supercomputers. Their use in this regard is important for stimulating formulation of algorithms and developing software approaches that are not practically tested on super minicomputers.

Recommendations to the National Science Foundation.

1. *Inform life scientists* of the NSF supercomputer initiative.

2. *Encourage immediate application of Class VI computers* in areas of research where present communication capabilities are adequate.

3. *Provide support for training investigators* on utilization at supercomputer sites and encourage training in vectorized software approaches and computational methodologies.

4. *Establish and support consulting* for user groups to aid in exchange of information on effective advanced computing.

5. *Establish telecommunication links* between supercomputer sites and investigator institutions that have a capabilities of 9600 to over 56K baud data rate transfer.

6. *Provide support for graphics capabilities* at user sites that are compatible with the telecommunications developed.

7. *Encourage some supercomputer facilities to acquire specialized devices* such as high resolution graphics and hard copy, on-line laboratory environments for real time experimentation, and data base storage for access at high rates.

8. *Ensure funding for supercomputers from sources other than the existing programs in the life sciences* where research support even now falls short of adequate levels.

9. *Stimulate cooperation between federal funding agencies* (notably NIH and NSF) in providing access to supercomputer facilities and in support of ancillary developments such as telecommunications and new software.

Specific Comments.

Many of the recommendations listed above transcend disciplines and bear directly on the success of the supercomputer initiative. The specific needs for considerations are discussed below.

Informing Potential Users.

It is important that all potentially interested researchers are informed of the supercomputer initiative, training, potential applications and means of applying for available funds and allotted cycle time. Also the scientific community should be intimately involved in the shaping of the NSF policies which attempt to address our concerns.

Training.

Life scientists without supercomputer experience will require special training. Currently training is being offered to novice users at the sites which are providing time on supercomputers. *Therefore funds must be made available to provide for transportation and lodging for investigators to travel to these sites.*

Three or four day seminars at the sites dealing with issues such as job flow control, vectorization, communication protocols, input output procedures, and program optimization are essential. Other issues that should be addressed through training include selection of the most efficient algorithm for classes of numerical problems, and graphical presentation of data and computed results.

Consulting

Consultant availability is critical for the success of the supercomputer initiative. Twenty four hour access to on site consultants via telephone for routine problems such as program bugs, compiler errors, protocols and procedures is essential. There will also have to be consultants who can advise users on the availability of software from existing libraries, use of system utilities, and on the selection of optimal numerical techniques. Because the biological community in general has not had access to supercomputing in the past, the NSF should make a special effort to provide true expert consulting in numerical analysis. Moreover, it should provide resources to support special workshops and seminars at national meetings to raise the level of numerical and statistical expertise directly related to supercomputing in the biological community.

Communication

Very high band width communication links between the supercomputer sites and the universities where NSF supported research is carried out is essential. Many universities where NSF supported research is ongoing are installing local networks that are capable of high speed communication on campus. These networks should be linked to the NSF supercomputer centers so that a minimum of 9600 baud communication is standard. Additionally, the higher communication rates required for graphics or research projects that capture and need to transmit large data sets rapidly must also be made available.

Graphic Workstations

Supercomputers can produce large amounts of data that are best viewed graphically. The graphic workstations and broad band communication devices that can handle this information must be readily available at the investigator's site.

Source of Funds for Support of Supercomputer Centers.

Support for supercomputer centers should not be obtained from the already strained support for basic and biomedical sciences. The applications of supercomputer does not substitute for current computing capabilities but allows greater amount of time to carry out other tasks on laboratory computers.

Coordination and Cooperation of NSF and NIH Roles in Providing Access to Supercomputer Centers.

Many scientists work in domains of both NSF and NIH support. In other cases, entire laboratories have need for access serving both research support entities. In order to prevent duplication of expensive facilities, it is essential that a coordinated effort be implemented which can best serve the scientific community as a whole and for specific applications. Duplication of data base access and telecommunication will no doubt be the result unless a coherent plan for support of advanced computer applications in biology and medicine is instituted.

Workshop Design.

On December 10-12 1985, a workshop supported by NSF, through a grant to New York University, convened at the Airlie Foundation, Airlie Virginia, to discuss advanced computer applications and needs in the life sciences. The proceedings consisted of ten workshop groups discussing specific issues of their respective fields and three minisymposia to acquaint scientists with supercomputers and related problems of graphics and telecommunications. The focus was on barriers to advancement within life sciences disciplines and how they might be resolved by advanced computational capabilities. The small groups reassembled into three larger groups to discuss issues that transcended the individual research approaches. These findings were reported to the group as a whole with general discussions centered around common computational questions and issues of application in life sciences.

Issues raised by the small and medium sized groups were related to molecular data analysis of structure and to organizational dynamics of organelles, cells, neuronal circuits and organs. These and other groups addressed modeling of dynamic properties of life constituents, their movements, distributions, and interactions. The scope of the discussions ranged through the technical aspects of advanced computer hardware and software, the recording and analysis of biological data, the representation of these data in models and interactive displays necessary to evaluate and demonstrate structure and dynamical events.

The participants were representatives of specific aspects of fields that might have need for supercomputer time, and experienced users who were able to relate information on the limitations and potentials of this advanced technology. The common needs toward promoting utilization of supercomputer time and establishing effective access were addressed in general assemblies. The results of the meeting are set forth in conclusions, recommendations and reports representing four well defined application areas. The four groups represent natural division of the discussion groups and their interests. The program of the minisymposia, list of interest groups and participants also are provided below.

REPORTS ON APPLICATIONS IN SPECIFIC AREAS OF RESEARCH.

A. Biomolecular Structure and Dynamics in Biological Function.

Edited by Stephen Harrison and Wayne Hendrickson

Overview. The function of biologically important molecules depends on their conformation in three dimensions. We must determine, visualize and analyze three dimensional structures in order to understand mechanisms of enzymic catalysis, recognition of nucleic acids by proteins, interactions between cell surface molecules and their ligands, antibody/antigen binding, and many other dynamic events central to cell biology. The recombinant DNA revolution has dramatically expanded the nature of the questions that can be tackled and of the processes that can be understood. Expression of cloned genes and *in vitro* mutagenesis permit purification and modification of any protein or RNA for which a gene can be identified. The experimental methods of X-ray crystallography, electron microscopy and nuclear magnetic resonance are the most fruitful approaches for studying molecular structure in three dimensions. Theoretical methods for energy calculations and for macromolecular dynamics are important for understanding function and interactions. Quantum chemical calculations on ligands and substrates are significant for analyzing binding and reactivity. Predictions of the folded conformation of proteins and of RNA molecules, given their primary sequences, is a vital long-range objective, if we are to translate the vast and ever-growing number of known nucleic acids and amino acids sequences into a deep understanding of function.

Advanced computation is essential for both theoretical and experimental approaches. The major progress that can be expected during the next five years will only be realized through effective use of supercomputing power and through innovative exploitation of other advanced computing technologies. Anticipated landmarks include: *1) Determination and refinement of crystal structures showing protein/nucleic-acid interaction, antibody/antigen complexes, cell surface receptors, cytoskeletal elements, and membrane proteins. 2) Complete determination of small protein structures in solution using 2-D Nuclear Magnetic Resonance. 3) Automatic modeling of a protein structure, using constraints provided by homology to a known structure.*

1. Determination of Macromolecular Structures by X-ray Crystallography.

The basic problem in crystallographic analysis of macromolecules is to transform the many thousands of intensity measurements from an X-ray diffraction pattern into accurate atomic parameters for the crystalline molecule. The resulting structural images profoundly influence how we approach fundamental questions in many areas of biology. Numerous significant macromolecular structures remain to be determined, but unfortunately the structure determination process has typically been slow. Supercomputing power, coupled with recent advances in instrumentation and new methods, now offer the possibility for substantial enhancement in the speed with which structures can be analyzed. Two aspects of the analysis are particularly receptive to advanced computing - the phase problem and model refinement. Other aspects also require intensive calculation, especially for larger structures.

a. The phase problem. The central step in the determination of a protein crystal structure consists in obtaining an electron-density map from the measured diffracton intensities; or equivalently, in solving the phase problem. This is traditionally carried out by heavy-atom substitution methods: isomorphous replacement and anomalous scattering. Frequently, however, these methods fail to produce phases of sufficient quality. For such circumstances, any computational procedure capable of completing the phase determination

provides an invaluable tool to the crystallographer.

A now well-established procedure takes advantage of non-crystallographic symmetry elements often encountered in protein crystals, and of the absence of features from solvent regions. These calculations are heavy (100 hours of VAX 11/780 CPU time for a virus structure) and supercomputer power is undoubtedly desirable for solving the very large structures now of interest for cell biology.

A new category of methods, of a statistical nature, are currently being investigated. They are expected soon to provide a powerful adjunct to substitution methods, and possibly to supersede them altogether in the future (G.Bricogne, *Acta Cryst. A 40*:410-445, 1984). These computations are complex (e.g. 10,000 hours of VAX 11/780 CPU time). They lend themselves to vectorization, but they involve a tree-directed optimization that requires frequent man-machine interaction via a graphics terminal. The availability of supercomputer power will be necessary to allow the further development of these methods and to make full-scale computations feasible. The possibility of performing such calculations will play a major role in encouraging development of this software. This will ultimately speed up the process of protein structure determination.

b. Structure refinements. The final stages in structure determination involve refinement of atomic parameters against diffraction data. Accurate models are essential for interpretation of functional properties at the atomic level. For a 50K Dalton protein, the iterative refinement requires about 8 hours of CPU time per cycle on a VAX 11/780, and 50 to 60 cycles are required to complete the optimization. For a large molecular complex, such as a virus, the refinement is only possible on a supercomputer. For example, 65 cycles of least squares refinement of a small RNA virus used over 1000 hours of CPU time on a Cyber 205 (A. Silva & M. G. Rossman, *Acta Cryst. A 41*:in Press). The structures, now emerging, of protein/DNA complexes and antigen/antibody complexes will require comparable computing power.

Other parts of crystallograhic structure determination (e.g. data processing for large structures) can also take advantage of supercomputing power. But the greatest impact will probably be in opening up entirely new approaches. The supercomputer will allow the use of vastly improved physical models of the protein as a restraint in refinement - one in which dynamical properties are simulated in the calculation. Interactive refinement procedures should also be very effective. Finally, as the data base of three dimensional structures grows, so too does the importance of demanding comparative searches in these structures.

A concluding comment on the character of supercomputing needs in crystallography seems appropriate. Some of our needs are for "production runs", such as refinements, already carried out on supercomputers. However, others such as the statistical methods for phase evaluation, are developmental in character. It proves most effective to explore diverse options without special concern for optimization in establishing new procedures. For large scale problems supercomputers become essential for this development. Finally, enhanced computing power will enable certain problems, now handled in batch mode, to become interactive.

2 NMR Determination of Macromolecular Structure.

NMR provides unique information as a result of the known geometry-dependence of relaxation times and of nuclear Overhauser experiments, which yield interproton distance measurements of up to about 4 A with reasonable precision. These measurements can be made in solution under varied conditions with motional and kinetic information also being

obtained. An important computation-intensitive application is the NMR determination of entire protein structures using methods currently being developed in a number of laboratories. This requires large-scale computation to analyze the data and to carry out the distance - geometry procedure. Current work suggests that the analysis time per experiment is several hours on a VAX, while the distance-geometry algorithm needs about 10 hr/structure for a 5 KD protein. Since many trial structures are generated to test uniqueness, the total computing time can exceed *100 hours/experiment*, whereas data collection requires 24 hours of VAX time. This experiment is now applicable to low molecular weight (approx. 5 KD) proteins as well to DNA duplexes of about 10 base pair. It may be extendable to 10-12 KD proteins. The general method is also applicable to ligands bound to macromolecules and to exploration of active-site geometry. Genetic engineering enhances the power of this approach. Supercomputer access would reduce the computational burden to 1-2 hours per experiment, and interactive use would be valuble as well.

Initial analysis of NMR data is generally carried out on-line, and VAX computers (or smaller) are adequate for such use. In large-scale applications, however, especially 2D NMR of biochemically labile preparations, it will be desirable to have access to supercomputing to analyze preliminary data sets, in order to determine the future course of a time-limited experiment. Such analysis might include currently-developing statistical analysis of spectra using entropy-maximization methods, which involve lengthy calculations.

Future magnetic resonance applications may be line-shape and relaxation analysis of ESR, deuterium NMR, and similar spectroscopies, using Monte-Carlo simulation of motion with many trial parameters to fit line-shapes.

Programs to calculate spectral features from magnetic resonance data, as a function of structure and other parameters, should be more readily available. These should be coupled to data banks of molecular structure and molecular motion characteristics. Access to this information should be through suitable display systems.

3 Theoretical Approaches and Molecular Simulation.

Computational models for macromolecules conformation and interaction are greatly limited by the power and character of existing computers. Supercomputers and innovative special purpose computers will alleviate current limitations and make possible qualitatively new developments, such as adequate treatment of solvation effects. Three important projects can be described that will benefit from substantial increases in computing power and speed.

a) Simulation of large macromolecular systems in as realistic a way as possible. Such a simulation will, in general, include thousands of water molecules in a periodic system, and it will use the method of molecular dynamics to model the behavior for 100's of psec. Such computations demand thousands of hours of VAX 11/780 time.

b) Automatic modeling of proteins using constraints provided by homology to a known structure. This is a particularly exciting goal, in view of the explosive growth of sequence data. Structural homology groups are starting to be recognizable, and attempts can be made to model important members of these groups from known structures having a related homology. Once the starting conformation has been defined, energy minimization is used to give a stereochemically acceptable model. Different starting conformations give a family of possible conformations, and molecular dynamics can be used with additional minimization to get to a lower energy. This same method might yield atomic coordinates that are guided by information on inter-proton distances obtained from NMR. These computations demand tens of VAX 11/780 CPU hours and the possibility of nearly interactive response with

supercomputers is an exciting prospect.

c) Prediction of the properties - especially the stability - of mutationally altered proteins, in relation to the properties of the starting structure. A facility in such prediction is especially important in designing enzymes and other proteins with altered characteristics. Studies of variants from *in vitro* mutagenesis can give a large data base for such computations, which can be used in turn to adjust the potential functions. Methods will be used that are similar to those described for homology modeling, and the computational demands are comparable.

The quality of a molecular-dynamics calculation ultimately rests on the intermolecular potential functions used to compute configurational forces and energies. At present diverse prescriptions are available, but there is major need for more extensive characterization and evaluation of these potentials with respect to experimental data. Thorough study of even a single system is a formidable computational effort utilizing thousands of VAX 11/780 CPU hours.

We note that it is now possible to calculate free energy, the fundamental index of thermodynamic stability. Methods that yield free energy require an order of magnitude more calculations than normal simulations. Such detailed computations will make macromolecular simulations more realistic and more informative.

Simulation of molecular properties can usefully benefit from the computational methods of quantum chemistry. Quantum calculations are important for establishing potentials used in modelling simulation. The reactivity of substrates and the chemical properties of ligands can also be computed. Optimal methods exist to tackle such problems correctly, but computing demands are substantial (thousands of VAX CPU hours). Supercomputers will allow extension of the molecular dynamic methods to model transitions in low energy, eg. quasi equilbrium states such as enzymic processes. Such transitions are important for enzyme catalysis and macromolecular conformational changes. We can foresee the combined use of quantum mechanics and molecular dynamics in study of macromolecular function.

4. Analysis of Primary Sequence Data and their Correlation with Structure and Function.

The "Gen Bank" genetic sequence database now includes sequences totaling about 4 x 10^6 bases; in less than 5 years, it will almost certainly exceed 10^7 bases. Amino acid sequences are collected in the database of the Protein Identification Resource. The extent of these databases, plus the necessarily complicated measures of relationship determinations which must be employed, means that database searches for sequence relationships are in general computationally intensive operations.

Algorithms have been developed that locate all sequences or sequence fragments within the database whose degree of relativeness to a particular sequence satisfies a well-defined criterion. Current application of this optimally vectorized supercomputer search to a sequence of a few thousand bases in relation to the present Gen Bank database takes several hours of CRAY I time. The extent of the search depends on the the measure of relatedness that is used. We anticipate an overall demand of several thousands of supercomputer hours over the next few years in the investigation of sequence relatedness. In addition, for the specific goal of identifying protein structural homologies, more sophisticated avenues are being explored. These methods will attempt to integrate other information, using current and projected developments in the areas of pattern recognition and expert systems design.

5 Molecular Graphics: An Essential Tool for Analysis and Communication of Molecular-Macromolecular Structure.

Interactive computer graphics is an essential tool in current macromolecular structure determination, analysis and simulation. Thus as supercomputation becomes more accessible, it is important that it should do so in an environment that supports the interactive graphics process. Capability of such man-machine interactions will enable a qualitative leap from batch analysis to interactive processing of intensive structural calculations. This approach will play a critical role in speeding the flow of investigations by reducing the turn around time from days and weeks to minutes and hours. These temporal dimensions are within the time frame for efficient flow of human thought processes that determine the subsequent analysis.

The key component for evaluating a graphically interactive mode is the communication rate to and from the supercomputer. Typically the interactive session will drive the submission of computationally intensive work to the supercomputer, furnishing the current state of the model as data for that computation. Results from the computation will then be transfered back to the local site, providing updated information for the working model and initiating the next round of local graphical interaction. For this mode of operation to be effective, results on the order of a megabit must be transmitted within a minute, so that ready access to at least a 56K bit/sec line will be essential for an operator. It will also be vital to maintain and upgrade local VAX-type computing strength, in order that graphics be effective. Indeed, we emphasize that optimal use of supercomputing units will only be possible if local computing capabilities are maintained at the highest level.

B. Mathematical Modeling of Biological Systems on Supercomputers.

Edited by Charles Peskin

Overview. For those modeling biological systems on computers, the NSF Supercomputers initiative is a most welcome development. There is a urgent need for more computational power, and modelers of biological systems are in an excellent position to take advantage of it.

A common theme is that supercomputer access will make it possible to use models that are far more realistic. In many cases, we already know what the appropriate equations are, but we have been unable to solve these equations in a realistic setting because of lack of computer power and readily addressable memory arrays that are large enough to develop the analysis. Supercomputer access will overcome these unrealistic symmetry restrictions, thereby increasing the number of spatial dimensions from one to two or from two to three, depending on the application. This will allow generation of models that truly represent the complexity of the organism. Because our work falls in the category of number-crunching, supercomputers are the appropriate tool.

Two issues that deserve further consideration are communication and training. Many biologists need to transmit large amounts of data at high speed to and from their computer. Even modellers like to work in an interactive environment so that they can quickly explore the consequences of changing parameters. Training is important because at the present time there is only a small group of biologists with enough mathematical or computational background to take full advantage of this new resource.

1. 3-Dimensional Analysis of Dynamic Processes within Complex Geometrical Spaces.

a. Blood Flow. An important use of supercomputers will be the numerical solution of the three-dimensional Navier-Stokes equations in the presence of complicated, moving elastic or muscular boundaries. Some examples are blood flow in the heart, secondary flow patterns in curved and branching blood vessels, and the interaction of fluid dynamics with blood clotting. Such studies are important for the design of prosthetic devices, and they also provide a means of studying normal physiology and disease processes in the computer.

While numerical methods for fluid dynamics vectorize very easily, it is important that the treatment of the boundary conditions does not spoil the vectorization. One solution of this problem is the representation of the boundary in terms of a system of forces, as in Peskin's work on blood flow in the heart and in Fogelson's work on blood clotting.

b. Mechanism of Cell Locomotion and Tissue Deformation. Differential equation models characterize the forces a cell can exert upon other cells or other substrates when its cytoskeleton contracts or relaxes. The contractile state of the protein machinery in the cytoskeleton is modulated by various "trigger" chemicals, principally calcium ion. The chemical kinetics of these trigger chemicals are influenced, sometimes dramatically, by geometric-deformational states of the cell.

Because cell motions occur very slowly, inertial effects are nil. Viscoelastic effects dominate. A serious side effect is that the differential equation system becomes implicit in form:

$$A(Z)\, dZ/dt = F(Z,t)$$

where Z is an n-vector describing the geometry of the tissue and its chemical state and A is a (banded) $n \times n$ matrix whose components are non-linear functions of Z. To compute dZ/dt at each time, vast linear algebra calculations are essential. That is, the hardest part of the calculation is already vectorized.

By coupling together many copies of differential equations, a model of the mechanochemical state of single cells can be ascribed to a mathematical model of tissue composed of many cells. This tissue might be the epithelium constituting an early embryo, eg. the mechanism of optic vesicle formation etc. For two-dimensional tissue models (cell ribbon caricatures), there are typically many hundreds of differential equations, but the model's dynamics can be computed without supercomputers - in a day or so of VAX 11/780 CPU time. To make such models biologically accurate, extension to 3-Dimensional simulations is essential. Then the number of differential equations increases by tens of thousands along with the complexity of each equation. Supercomputer number-crunching prowess becomes essential.

c. Reaction Diffusion Models in Morphogenesis. The formation of models from reaction-diffusion patterns, as initiated by Turing in 1950 and later perfected by many biomathematicians, require vast computational resources when extended to three-dimensional models. The number-crunching task is in the 'solution of several coupled non-linear quasi-linear parabolic PDE's. A subsidiary task is "recognition" with appropriate graphics display of constant-concentration loci at each of many instants as the solution evolves. Software for both tasks is available and is easily vectorized. The constant concentration loci locator algorithm is inherently split into many parallel performances of the same generic task.

Mathematically identical, but biologically distinct, are problems of understanding how waves of depolarizing action potentials are propagated through two and three dimensional domains where there are partitions of electrically excitable membrane facets. Neural nets and myocardium are two examples.

2 Movement of Molecules and Particles in Biological Media.

a. Diffusion of Ions. The electrical activity of nerve and muscle cells is produced by ion transport through cell membranes and is determined by the electrical potential due to the local concentrations of ions and other substrates. Theoretical studies to date have concentrated on spatial variation in electrical potential, but there is now sufficient experimental evidence to show the importance of contributions from variations in ion concentrations in the restricted spaces around cells. The solution of partial differential equations is required to predict such variations with one equation for each ion. Only a few calculations of this type, to date, have applied to the single dimension analyzing up to three ions. This requires many hours of VAX time per run. More realistic problems for disease-related drug studies will require expansion to two or three dimensions, placing the computations out of the range of conventional computers.

b. Transport of Solutes through Porous Media: The transport of water and substrate molecules across the clefts between endothelial cells and through matrices of interstitial or intracellular structures is describable in purely hydrodynamic terms in relatively simplified situations. These describe solute geometry, torque, local pressure gradients, shear, viscocity and the geometry of channel walls or matrix. A general method for describing movement of isolated solute molecules now needs to be extended to heterogeneous fibre matrices. The results should be applicable to describing filtration, diffusion, volume exclusion, solute and solvent drag, and in general, both the fluxes and the phenomenological coefficients summarizing the behavior of the system.

c. Fate of Toxic Materials and Drugs. Prediction of the effects of drugs and toxic materials in animals and humans requires a knowledge of the distribution of the substance in the body or organ so that local dosages are accurately represented. Calculations of such distributions involve the solutions of fluid flow or diffusion equations in two or three dimensions with boundary conditions from other processes such as absorption, breakdown, or transport. Present models use a small number of compartments or one-dimensional transport as a rough approximation. Supercomputer power is needed to perform more realistic calculations.

d. Brownian Dynamics. Brownian dynamics simulates the motion of ions in a channel by a combination of Newton's laws, potential terms describing the channels, and a stochastic fluctuating potential. Because time steps as long as 15 psec can be used, it is feasible to integrate through the physiological time scale of milliseconds, on a supercomputer. The advantages of molecular dynamics is that it is within the range of modern computation. The motions of the ions are computed, the channel protein is represented by a potential energy barrier with water being represented simply as a frictional term. A Langevin equation is written to describe the ionic motions, in which a stochastic driving term is added to the potential functions and frictional terms described. This equation can be integrated in minutes of CRAY-1 time (for milliseconds of real time). Such simulations will allow for more realistic modeling of the ion permeation processes which underlie a large number of biological phenomena. Since measurements of single channels are now routine, a productive interplay of theory and experiments can be anticipated.

It can be reiterated that these computations will be greatly aided by substantial access to time on supercomputer, but ONLY if easy access from the laboratory with a 50K baud/sec data rate is available using standarized graphics and alphanumeric protocols.

3 Mathematical Modeling of Initiation and Propagation of Electrical Activity.

a. Ion Channels and Pharmacological Interactions. Access to single channel events provides an important tool for viewing first hand, the result of conformational changes in single protein structures - i.e. channels. These processes are stochastic in nature. Transition probabilities between conformational states can be determined from direct measurement of ion flow within the channel.

Pharmacologic agents can modify channel conductivity. For instance, some ion channel blocking agents appear to diffuse into the channel core during the channel open time and bind to some interior surface feature resulting in an occluded channel. During the channel close time and under appropriate conditions, the drug can diffuse away from the channel, but is unable to diffuse back into the channel until it opens again.

b. Electrophysiology - Heart. The heart is a network of interconnected excitable cells that under normal circumstances beats with a periodic rhythm. Under pathologic conditions, disruption of cardiac rhythm occurs, and this can frequently be controlled with ion channel blocking agents. The mechanism of action is just beginning to be understood. The major battleneck in investigating antiarrhythmic drugs is in scaling up descriptions of single channel blockade to multiple channels in a cell and to multiple interconnected cells. Several alternatives for describing such complex cellular arrays are available. One approach is through use of parabolic PDE's to describe propagation and blockade along a cell and coupling patterns of individual cells.

Mathematical models have been developed to investigate the ionic bases of the electrical activity of single cardiac cells, including impulse initiation and propagation of the wavefront.

Recent models are based on Hodgkin and Huxley-type equations that describe time- and voltage-dependent ionic currents. These models have been used to mimic successfully the electrical activity observed experimentally in a variety of cardiac cell types. Attempts have begun to generate one- and two-dimensional networks of these simulated cells. Results of these multicellular models have provided new insights into the mechanisms of pacemaker synchronization, wavefront propagation and various cardiac rate and rhythm disturbances.

However, it is obvious to investigators in this field that currents are greatly oversimplifed and that access to supercomputers is essential for the development of more realistic models that more closely approximate the biological conditions. For example, with greater computer processing power and the capacity to handle much larger data arrays, individual cells can be modeled in three-dimensions for their cellular compartmentalization and metabolism in conjunction with integration through ionic mechanisms. Similarly, the increase in computer power would permit the development of three-dimensional models of networks of simulated cells. It is anticipated that these more powerful and realistic models will contribute importantly to our understanding of normal and abnormal cardiac rate and rhythm and to neuronal function.

Models of the cardiac action potential are typical of excitable membrane models and include the effects of numerous ionic channels at the cellular level, and coupling between cells to induce propagation. Although the electrical phenomena obviously determine the mechanical and fluid dynamical function of the heart, the importance of mechanical feedback is not known and is not yet included in state of the art models. Similar modeling problems occur in smooth and skeletal muscles where the propagation of electrical signals is again responsible for proper function.

Simulation of these models is currently done only on a small scale. There are significant numerical difficulties associated with the stiffness of these systems, which are amplified when they are distributed in a spatially inhomogeneous medium. Accurate, stable computation with only modest spatial refinement certainly requires supercomputer power.

Another scheme is to define a finite state machine for each cell in the network that represents the appropriate submicroscope elements of electrical activation and channel blockade. An array of these elements becomes a cellular automata, from which macroscopic behavior can be assessed. Use of this strategy, though, depends on careful "calibration" of the submicroscopic state in reference to actual cellular behavior.

Supercomputing can play an important role in addressing questions of multicellular electrical activation as modified by ion channel blocking agents. High resolution displays are needed to visualize patterns of cellular activation.

4 Modeling of Neurons and Neuronal Circuitry.

Cellular Neurophysiology is one of the more quantitatively developed life sciences. A few experimental model systems have led to the development of biophysical - mathematical models for the generation and propagation of nerve impulses, the self-sustained oscillations of pacemakers, and the electrical activity throughout the branching of a dendritic neuron. These models involve linear and nonlinear partial differential equations and require seconds to minutes of computation time (e.g. DEC-10) for their solution. These models have played a key factor in how we understand integration and signalling in neurons.

As experimental techniques improve and more complex neuronal systems are studied, we find that many neurons and excitable tissues do not satisfy the assumptions of the classical

models. Neuroscientists are seeking to relax some of these assumptions (such as distribution of ion channels over cell surface, types of ionic channels, idealized branching geometry, one-dimensional cable theory) and to compare, quantitatively, theory with experiment. Such extensions place increased demands on present computational approaches. For example, a five millisecond simulation of the voltage transient in a neuronal dendrite with several hundred branch segments may require tens of hours on a VAX 11/750.

Moreover, such simplifying assumptions as uniform and constant ionic environments are inapplicable in a number of cases. Incorporation of the three-dimensional aspect of intra- and intercellular diffusion, binding and uptake will certainly tax the capabilities of supercomputers. Applications of these enhancements to synaptic transmission, ion-activated channels in excitable membranes (e.g. Ca^{++}-activated K^{+}-channel), are being found in many neuronal and secretory systems and can lead to complex signalling such as bursting, excitation-contraction coupling in skeletal muscle, and spreading depression.

Biophysical models for individual cell behavior are beginning to be applied to investigate multicellular/network phenomena. For example, various mechanisms for cellular coupling (electrotonic, chemical, ephaptic) and its role in epilepsy are being considered with computational models of arrays of dendritic neurons.

Intercellular connections are being mapped from simultaneous recordings of two or more neurons in relatively small functional networks of invertebrates (e.g. motor pattern generation). These data will likely soon be incorporated into computationally intensive mathematical models. Other applications of multicellular modeling are for synchronization of electrical activity in some secretory systems and for sensory processing (e.g. cochlear processing involves three-dimensional fluid dynamics, mechanics and electrical signal generation in hair cells with synaptic transmission to auditory nerve fibers.

5 Population Dynamics in Biological Systems.

a. Spatial Heterogeneity in Insect Ecosystems. Successful biocontrol of agricultural insect pests using appropriate insect predators (not insecticides) requires quantitatively accurate models with definitive predictive power that can account for spatially localized outbreaks in parts of the system and the spatial migrations of predators to cluster at and decimate pest infestations before they spread. Such models, recently derived (cf. Kareiva and Odell) in the form of quasi-linear parabolic PDE's, appear able to do this. Predator chemotaxis is a central aspect of the model from observation of predation/reproduction cycles of individual insects existing together in cages. This information is used in the PDE models to mathematically derive the outcome.

The numerical solution of such systems (two spatial dimensions are essential) can predict, from initial conditions on insect distribution, how a spatially heterogeneous insect predator-prey ecosystem will evolve. While parabolic, the PDE system conducts wave-like solutions in which shallow gradients in population density are transformed into very steep gradients. Substantial number crunching capability is required to solve these systems. Supercomputer prowess would make it possible to predict the evolution of pest outbreak and control before they happen so that corrective measures (planned release of additional predators where needed) could be taken in time.

C. Analysis of Dynamics of Biological Processes.

Edited by Rodolfo Llinas

Overview. In the most general of terms biological processes represent a set of spatio-temporal domains of function which, for their proper study, require extensive numerical computation and data assembly. From this point of view, computers are essential in dealing with physiological complexity, both from a theoretical modelling vantage-point as well as in dealing with the massive amounts of data which can now be gathered in these types of studies. Only recently, as a result of computer use, has a realistic analysis of multiple variables in the temporal domain been possible. Because the range of studies which ultimately relate to the analysis of dynamic biological processes extend from the molecular to the social, the utility of supercomputers must be considered as no less than crucial.

1 Mechanisms of Membrane Dynamics.

All animal cells are defined by a membrane which separates their cytoplasm from the extracellular milieu. Research on membranes has been a central theme in biology for nearly a century and has afforded a true conceptual and experimental evolution. For technical reasons membrane ensembles do not lend themselves to direct molecular analysis. Thus, a large background of analytical detail was necessary before the more synthetic type of measurements could be interpreted in terms of the molecules which control biological functions. Thus, this integrative area of research has been rather isolated until recently. Indeed, the field of membrane biology is just beginning an era of growth.

a. Channels. Most biological functions of membranes are mediated by specialized glycoproteins inserted through the membrane. These "channels" provide the only gatable pathways for natural transport across the membrane. The macromolecules which generate such channels (and to a lesser extent the membrane structures themselves) have evolved, becoming adapted to meet the needs of survival.

b. Statistical Analysis of Ion-Channel Data. Since 1980 the technique of recording from single ionic channels has spread explosively. This powerful experimental technique allows a researcher to study the behavior of a single transmembrane protein molecule for many hours in its natural environment. The measured behavior is the current (appr. 1 picoampere) which turns on and off randomly as the channel opens and closes. This behavior known as gating is rich with kinetic information pertaining to the underlying molecular properties of the channel, but because of the inherent randomness, this information is quite difficult to extract. In a typical experiment more then 5 megabytes of information is acquired from a single channel for later analysis. The principal roadblock to an understanding of the molecular mechanisms underlying the gating of ionic channels is the analysis of these data.

In the past year powerful methods have been developed for such data analysis, based on maximum likelihood techniques. These methods are not feasible for typical situations, even with full time use of VAX computers, for the computer time to analyze a single experiment can be more than one month. Fortunately, the problem is easily vectorized (see, e.g., Horn and Vandenberg, 1984), and is therefore ideally handled by a supercomputer, where such calculations can conceivably take from a few minutes to hours to compute. The availability of supercomputers to the many researchers studying single ion channels will make a substantial improvement in the now-Herculean task of data analysis.

2 Investigation of Distributed Information Processing and Control in the Brain.

a. Computational Requirements. When the brain and the computer are compared, it is the contrasts rather than the similarities that stand out most prominently. To the extent that we understand the operation of the brain, we can say that it is based on the continuous, simultaneous operation of a large number of differentiated neuronal assemblies that are heavily interconnected by tracts of parallel coursing fibers. The challenge of this field is a) to understand the functional block diagram of the nervous system, b) to understand how control signals and information are represented by the patterns of neural activity transmitted over the parallel fiber tracts, and c) to understand the operations of the neuron assemblies.

Progress in this field has been greatest near the input (sensory) and output (motor) sides of the nervous system. Relatively little is known about the mechanisms of the intervening processes. Most of the current progress in the field has been based on electrophysiological recordings from single neurons within the neuronal assemblies or single nerve fibers within the tracts. There is now widespread agreement that, because of the parallel nature of processing within and between modules, multiple simultaneous recording procedures are essential for each of the objectives referred to above.

The specific forms of data analysis in this field are not well developed, but their general form can perhaps be inferred from the objectives. The "block diagram problem" will involve the analysis of coincidences and dependencies between activity patterns in the modules and between this activity and the animal's behavior. The "neural image" problem (the form of information representation in parallel pathways) will most likely involve analyses of an image processing type. The "neural network" problem (operations within a single module) will involve methods aimed at inferring the functional organization within a neuronal assembly. Current efforts in this area center on cross-correlation methods.

b. Analysis of Spike Trains for Determination of Intra- and Interneuronal Integration. A form of analysis that can be predicted with relative certainty is the use of simulation as a part of hypothesis testing. This type of analysis is noted for its heavy computational load, and it is compounded by the extensive data in this field. The data load imposed by such experimentation may be estimated by considering the nature of the experiments being planned at present.

The experimental design involves a primate trained to perform a sensory, motor and/or cognitive task. Then, multiple electrodes are inserted at specific brain locations depending on the system or systems being examined. Currently, the number of electrodes is 10-30. The aim for the near future is to increase the order of magnitude by one. During these experiments, the animal is subjected to a wide range of experimental conditions (e.g. a large number of sensory patterns while performing a pattern discriminating task). A single day's experimentation will generate several hours worth of data. The mean impulse rate from each recording site is in the order of 100 Hz (short term rates range from 1 to 500 Hz). From these figures it can be seen that data rates are of the order of 100,000 impulses per recording site per hour of experimentation. Modest numbers of recording sites generate data rates of the order of 1,000,000 impulses per hour. One hour represents a relatively brief period of behavioral experimentation.

This field is not sufficiently well developed to provide an accurate specification of its computational requirements. However, multineuronal recording is rapidly being taken up by laboratories around the world and it is clear that the computational load imposed by the data and analysis methods is a major problem. What is needed in the near future is a flexible basis for exploratory data analysis and display such as that provided by the "S" package developed at Bell Labs. Supercomputers are required for the large data volumes generated in

this field. 0.1.21

c. Auto Correlation and Cross Correlation of Data Generated from Temporal Analysis of Multiple Neurons in Circuits. A significant extension of data analysis has been achieved with GACF, the Generalized Auto-Correlated Function (T.J. Ebner and J.R. Bloedel: *J. Neurophysiol. 45*:5, 1981). This technique pertains to analysis of spike trains from a single cell and has been generalized to the Generalized Cross-Correlation Function (GCCF), for understanding the integration of activity in neuronal populations.

The GACF outlines the effect of a stimulus on the auto-correlation of a spike train at various time lags. If 1000 time lags are chosen for both the stimulus and the auto-correlation, then each spike in the train would require about 1000 calculations. For example, if one examines 10 spikes in 1000 time frames after the stimulus, there are about 10,000 computations per stimulus.

We generalized to the cross-correlation of each cell with each other. In general, that means:

$$n!/k!(n-k)!$$

possible correlation sets where n is the number of cells and k is the number that are used for comparison. Thus if GCCF is applied between any two cells there are:

$$n!/2(n-2)! \text{ or } n(n-1)/2$$

combinations. For example with 100 cells, this is about 5000. If each GCCF takes about 10,000 computations, then a complete set would involve about 50,000,000 computation sets. The data is in the form of an array of n elements for each time frame (chosen here as 1000). There is a new array for each stimulus. (The technique used on the VAX has an n of about 2,000,000 array and the n x 1000 arrays are actually subsets of the large array. Ultimately, the array will best be an infinite stream of samples).

Given a stimulus at one Hertz we have about 50 million computations per second for cells taken 2 at a time. Taken 3 at a time, a group of 100 cells increases the number of computations by a factor of 32 to 1,500 million computations per stimulus. Multiply this by the number of seconds (or stimuli) needed (perhaps 500-1000 or more).

In some experiments it would be required to sample the data at a number of time lags after the delivery of a drug. In addition each of the GCCF's requires a 3-dimensional surface rendering for visualization. This is also a tremendous computational load. The 3-dimensional hidden-line surface projection of a 1000 x 1000 matrix would also be required.

3 Population Genetics.

Three problems require advanced computers: macromolecular homology, phylogenetic trees and pedigree likelihoods. Other areas make increasing use of computers and may develop a need for extremely fast computation in the near future. These applications are CPU intensive with modest I/O requirements. There is little demand for high speed data transfer, real-time computing or image generation. On the other hand, there will be an urgent need for algorithms in numerical analysis to provide function minimization that is parsimonious for vector cycles, unlike conventional methods that minimize the number of sequential function evaluations. These computer bound applications seem ideally suited to supercomputers.

Looking for homologies of nuclei acid and amino acid sequences between loci and organisms is a major task of evolutionary biology. Initially, this may be a simple search to see if any part of a sequence matches any part of the national library. A second level is the finding of an optimal global alignment between two sequences with computer time usually of order N^2 , although Nlog(N) is possible. A third level is the simultaneous optimal alignment of K sequences which is currently of order N^K. Improved solution of this problem is important for studying the evolutionary relationships among distantly related sequences and for finding consensus sequences that might reveal functional subregions.

Much of evolutionary biology is concerned with the inference of phylogenetic relationships (populations, species, proteins, or nuclei acid sequences). This requires fitting the data to a tree (cladogram) and seeking the tree to which the data best fit. There are many kinds of data, many methods of finding appropriate trees and many criteria for what constitutes the best fit. The number of possible trees for even a moderate number of taxa is so large that even a supercomputer could not elaborate them all, but obviously the reliability of the conclusions increases with the number of trees being explored. The number of possible tip-labeled unrooted trees for taxa is $1*3*5...*(2t-5)$. For only 21 taxa this is greater than a "mole" of trees ($6*10^{23}$). Currently exploration of all trees is not feasible for $t > 9$. For each set of taxa there may be anywhere from 50 to several hundred items of information (characters, genes, morphological attributes, amino acids, nucleotides, etc). Each of these items must be manipulated (2t-1) times in a given tree. Thus for $t = 10$ there are $1*3*5...*15 = 2{,}027{,}025$ trees $*19$ manipulations. For only 100 items of information there are $3.8*10^9$ operations. Each new taxon would increase that number by more than one order of magnitude so that the task is clearly not feasible on ordinary computers and will remain so even in the face of improvements of the branch and bound type algorithms. A similar degree of difficulty arises if the items of information are at a genetic distance.

Whereas the first two problems might not lend themselves to vectorization, pedigree likelihoods form vectors whose elements differ in one or more parameters. Each such vector should lead to improved parameter estimates (by an algorithm yet to be optimized) and to either computation of the next vector or acceptance of convergence to the maximum likelihood solution. This kernel is basic to segregation and linkage analysis and to the determination of recurrence risks. Models much more complicated than the ones currently being tested cannot be studied with ordinary computers in a feasible time and will therefore be neglected until faster computers are available. Such models include genetic modifiers, family environment, multiple alleles, complementation and other factors that may be essential to the understanding of genetic factors in families and populations.

In the last few years population genetics has split into evolutionary and epidemiological branches. The latter, interested in genetic effects in contemporary populations, involves a group of several hundred investigators who come together before and during the annual meeting of the American Society of Human Genetics. These workshops provide quality control for methods of analysis and a forum to discuss research problems. Much of the effort has been devoted to the explanation of possibilities raised by fragment length polymorphisms, enough to provide close mapping of any DNA sequence. Efficient use of this expensive technology requires that a large number of polymorphisms be studied in the same families but it is not yet clear how loci can be studied simultaneously. Computing time goes up explosively with the number of loci and alleles, even if efficient algorithms to prune inadmissable genotypes are implemented. Solution of this problem determines the reliability of distance estimates, the power to exclude incorrect gene orders, and the prediction of multiple recombination fractions essential to genetic counseling and understanding recombination in man.

D. Image Processing and Image Generation.

Edited by Dean Hillman

Overview. Image processing and image generation research, whether related to life sciences or not, require massive computational power. Extended computation in image analysis has taken different approaches between fields. The crystallographers have chosen supercomputers to process and enhance recordings of molecular shape. Tomographic and microscopical image processing has not been addressed by Class VI computers but by specialized processors. These 2-dimensional frame pipeline processors can carry out limited functions on 2-dimensional arrays of pixels. Other processors are being generated to operate in the 3-dimensional domain (solids graphics engine, Phoenix Data Systems). These specialized processors, however, have limitations, and before elaborate designs can be generated, supercomputers can play an important role. The most advanced research applications, particularly in analysis of biological imagery, require processing that has flexibility in order to generate prototypes of new algorithms, non-standard procedures and architectures all being directed at fundamental problems of machine image processing and image generation. Development using supercomputers could be instrumental in overcoming barriers to experimental procedures where image analysis and image generation are essential.

The use of Class VI vector-type supercomputers in image generation is well established. Currently, the graphics film industry has made extensive use of these computers in producing various perspective views of static displays and for generating moving objects. The rotation of multiple objects and rendering their surfaces is computation-dependent. A number of procedures applied in image analysis and display must use the same computation-bound approaches as the film industry. Supercomputers will be very important in both image generation for model building and in methods requiring 3-dimensional image reconstruction for analysis of structural organization and dynamic processes.

Some image generation and image analysis applications can now make immediate use of supercomputer cycles given modest file transfer facilities. However, for most interactive image analysis tasks, off site processing is less well suited because of the limited access currently afforded by the NSF project. A much broader range of applications can be facilitated by a concerted effort to provide a unified approach to telecommunications and graphics display capabilities.

The sources of images in the life sciences are very diverse originating from variations in excitation sources and detecting devices. Light and electron microscopy and tomography are the principal images requiring advanced analysis at this time.

1. Image Processing of Light and Electron Optical Images.

Since the inventions of light and electron microscopes, morphologists have been uncovering the structural makeup of biological systems by observations made using preparations of tissue slices. By sectioning the object into slices that are thinner than the object, the surface of embedded structures are revealed in relation to the surrounding matrix. For the most part, this image domain has been considered to be 2-dimensional in nature, thereby avoiding many of the difficulties associated with 3-dimensional vision. Now with digital image processing, the handling of images has been improved significantly. However, neither the 2-D nor 3-D problems for low or high resolution analysis are solved since identification is easily confused by texture, incomplete boundaries, overlapping structures etc. These problems are compounded when extrapolating to the 3-dimensional world.

Cells are extraordinary complexes of interacting molecules, macromolecules and organelles that make up complex tissues of organs. In order to understand in detail how living cells and organs function, it will ultimately be necessary to elucidate not only the structure of these complexes but also the 3-D organization among cells. Optical microscopy, almost uniquely, allows cellular structures to be examined in the native state. Now with advancement in computational technology, quantitative analysis and 3-dimensional reconstruction have greatly extended the potential information that can be extracted from these approaches.

Ideally, analysis of macro- and microstructure can be made from 'three dimensional matrices of densities' represented as voxels (unit of digital volume). Digital methods are effective for separating embedded objects (having a defined density range) from surrounding objects in order to enable display of the image from various perspectives showing spatial relationships with surrounding and neighboring structures.

Microscopy remains as the preparation of choice for most applications where size, shape and relationships of internal structures are to be analyzed at high resolution. The resolution of tomographic instrumentation is limited and transmitted light microscopical approaches are not generally effective beyond 300 microns of specimen thickness.

There are four aspects of image processing that are separable but intimately related in a number of applications. These are: *a) image restoration, b) image enhancement, c) image reconstruction and display and d) image analysis*. Each part requires extensive computational power that is dependent on the type of operation and the size of the image. Currently the entire analysis process is restricted because of limits in the size of images that can be computed using specialized software and memory access of super minicomputers. Adequate algorithms that can effectively analyze images in an intelligent manner have not yet been determined. These limitations also apply to image generation techniques. Recent development of software requires powerful processors that have the flexibility to attempt involved approaches.

a. Restoration of images. Sectioning of preparations results in a need for *realigning* the objects to produce a reconstruction. The process of alignment becomes increasingly involved as the resolution demands are increased. This is most critical for recreation of the 3-dimensional density array of voxels from 2-dimensional pixel arrays. Additional processing is usually needed. The objects are frequently distorted differentially by the inherent forces of sectioning. Thus, image restoration as *warping* of the image back to the original shape is important as a means of reestablishing continuity as an array of densities. Alignment and warping to restore the 3-dimensional matrix must usually be done together and are very computation intensive processes since the third dimension involves correspondence determinations over a number of sections. Currently, hardware and software approaches are available for both procedures on 2-dimensional arrays of two adjacent sections. The extention of this to a number of serial sections increases the requirements for restoring the actual relationships.

Restoration of images from *optical sectioning* is a new method that promises to offer a quantitative, 3-D analysis of intact biological specimens by restoring the image quality acquired by the physical properties of the instrumentation. Essentially, the in-focus part of the image also contains out-of-focus components from the top and the bottom of the focal plane. Restoration of the in-focus image can be done in some naturally thin preparations (150-300 microns for light microscopy) or in slices containing the entire object. Here, focal plane sectioning is used to extract the voxel data in a digital form similar to the axial tomographic approaches. Optical sectioning eliminates having to warp and align sections.

However, this method can also be used effectively in combination with sectioning and alignment in order to generate a larger matrix in the third dimension.

In the optical sectioning method, 3-dimensional data is collected by a through-focus series of a complete specimen from one or more viewing directions. Each of the recorded images is a sum of in-focus terms from a narrow plane within the specimen and out-of-focus information from the remainder of the object. From a large set of planes, and a knowledge of the optical properties of the microscope, it is possible to remove most of the out-of-focus data from a single set of sections and all of the contamination outside of the object in question. The latter is subtracted by using two or more viewing angles.

Typically, each view of a data set consists of 64 to 128 frames of 512x512x8 pixels. Reconstruction requires a 3-dimensional deconvolution operation on this massive data set. Experience with related problems on smaller data sets suggests that the use of non-linear constraints significantly extend the resolution of the reconstruction. Iterative, constrained Fourier algorithms are particularly applicable.

Extracting the maximum amount of biological information requires that the best reconstruction be performed. Currently, this is a major research problem that cannot be explored due to insufficient computer time and memory. Therefore, supercomputers can play a significant role in initially testing alogrithm performance and perhaps later for routine data processing once a suitable approach is found.

Since the data is essentially visual, adequate evaluation of the results demands that the reconstruction be visualized on a suitable raster-display system. On-site calculation is possible but not desirable for the initial phases of this research. This approach only becomes practical if a display system with adequate resolution and reasonable transmission times is also present in the laboratory.

Remote access is made difficult by the problems of transferring massive amounts of data (25-100 megabytes of data) to the supercomputer site. This would only be practical by either high speed (1 megabit/sec) communications trunks or by mailing many tapes. The time delays caused by sending back tapes from each trial is unacceptable where many different trials are required.

In short, there is a significant and pressing need for supercomputers in the problem of optical sectioning microscopy. Unfortunately, the difficulties involved in data transmission may require that local solutions to the computational problem be found and this will take some time. For this application supercomputers are useful, but would only be compelling if a 1 megabit/sec transmission link could be established.

Restoration of images from *electron optical sources* is performed on the raw data to remove the effects of the instrument. These effects are due to geometrical distortions in the lens systems, such as spiral and pin cushion distortions and also to contrast transfer function effects. Distortions can be eliminated by mapping images onto corrected coordinate systems. Contrast transfer functions (CTF) are corrected by deconvolution of the image via Fourier transform methods. At present, these corrections are avoided by selecting small regions (all of which minimized the effects of distortions) or relatively low resolution (appr. 10A, where CTF corrections are less necessary). However, if high resolution (i.e. appr. 7A) is to be obtained from regular lattices of proteins, it will be necessary to employ large arrays (i.e. 2048 X 2048) and to correct data to make high resolution data available. At present, limitations are documented in these areas and have not been overcome due to the magnitude

of the computing problems.

Access to supercomputers would permit large arrays to be processed and provide data which could be used to obtain high resolution. The essential requirements are for interpolation and fast Fourier transform analysis of arrays of approximately 4000 x 4000. This will require vast computational time for multiple images and an adequate input/output communications system to the supercomputer.

b. Enhancement methods for light and electron microscopical images are related to visual optimization of density and color for interactive observation, and to noise reduction for digital processing. Processors capable of performing *image convolutions* at video rates are now in common use. These operations are effectively done in dedicated processors that shift the raster scan gray scale linearly or non-linearly in one or two dimensions resulting in contrast enhancement. This type of optimization is further enhanced by *color coding* of density levels (pseudocolor). The translation of gray scale to color coding acts to take advantage of gray level separation by visual perception of colors hues. Both of these procedures are currently available in hardware devices but much more advanced procedures are necessary to realize enhancements for noise reduction in large 2 and 3-D arrays. Supercomputers may be very beneficial in these operations.

Image enhancement in electron microscopy is aimed at *reduction of the signal to noise ratios*, and is done mainly by spatial averaging techniques. There are two main approaches: 1) averaging over several unit cells in a regular lattice of protein subunits. 2) averaging images of unit cells on a lattice of individual single isolated particles selected by correspondence analysis (CA). With the Fourier averaging method, a specific limitation is imposed by the time taken for the Fourier operation. In the case of the CA averaging method, the majority of computational time is taken in selecting particles and then sorting them into groups that fit images showing similar structural characteristics. In both cases, the benefit would acrue from the use of a supercomputer, particularly in the CA case where very large sparse matrices must be diagonalized.

It is important to view supercomputers as machines which will allow new processing schemes to be tested and evaluated. In addition, they will be important to bridge the needs in processing demands until optimal dedicated approaches can be generated. Supercomputer use should not be regarded as an end in itself. For example, developments in tomographic imaging obtained by use of supercomputer should lead to the development of imaging devices which can be made available to biologists and clinicians at reasonable cost. Thereby the development of specialized processors is an essential element of both basic and biomedical sciences.

c. 3-Dimensional Reconstruction and Display. Three-dimensional reconstruction is the generation of the surface of structures that are embedded internally. In the case of 3-dimensional arrays of data (voxel type), the extraction of dense objects and display of rotations have been coded in software and are now being produced as hardware operations. These operations are limited to a threshold of density for representing an object.

Alternatively, a method of preprocessing analysis can be used to produce low resolution reconstruction of images from high resolution data. This involves extraction of the boundaries from serial sections either before or after the aligning procedure. Automated boundary defining software is currently limited
to a threshold of density and therefore does not adequately determine the real
boundaryof the object. Interactive methods must be used. Supercomputers could be an important tool for defining algorithms that can do this tedious procedure.

The reconstruction of images from serial section microscopy is a semi-routine task within many biological areas. The images can be those that are preprocessed as described above. If they are obtained from sections, distortions usually make it necessary to warp the image in order to relax the effects of both the differential and overall compressions of sectioning. In light microscopy, both optical slicing and serial section aligning approaches can be used together providing that section alignment is relatively accurate. At the electron microscopical level, high resolution images result in intrinsic limitations making aligning of the density arrays very difficult. Supercomputers may be necessary to establish this offset because of non-linear distortions between sections. This step is currently not calculable with the small machines.

These sections are then used to yield composite images by rendering connected points to recreate the surface. One requirement for supercomputers will be in high resolution rendering of hidden surfaces obtained from reconstruction data.
The surfaces that extend between the boundary rings are connected and then rendered to produce lighting and perspective effects. Both procedures produce 3-dimensional images of objects that can be interactively viewed, selected, and deleted in order to uncomplicate the images.

High resolution display of lines, color and texture has been a standard feature of conventional methods of illustration of data which should be implemented for electronic based information. Rendering of surfaces by complex lighting algorithms involving ray tracing, shadowing, and texturing are practical on multiple images only with supercomputers. Input for high resolution imaging of complex three dimensional data structures can be done using very low cost personal computers for input but then this only can be processed effectively at a supercomputer center with an output record on film. A large capacity for high resolution image rendering using supercomputers is strongly encouraged as a part of the NSF supercomputer initiative.

The application of supercomputers to image generation has been developed and extensively used by the motion picture film industry and advertising interests. Efforts should be made to take advantage of this knowledge and technology making it available for the most beneficial solutions using graphic displays. The acquisition, manipulation and display of visual information in digital form places growing demands on supercomputer use. In the film industry, the limits are already encouraging construction of processors of new high speed design. These computers coupled with high-speed, high-volume data communication will be important tools in image generation from scientific data as well as mathematical models of biological processes. A desirable concept is a supercomputer center specializing in biological analysis would be a system of applicable software, a data base and graphics tools to enable high resolution display and graphics recording.

d. Image Analysis. Software for object recognition and quantitation has lagged seriously behind enhancement approaches. Consequently, except for a few application areas where image segmentation is facilitated by special tissue preparation (i.e. smear, suspensions, etc.), automatic image analysis is only poorly understood. This usually necessitates operator intervention to outline borders, exclude objects and identify structures.

The extraction of information from biological images is a monumental task. Generally, it can be partitioned into four operations: *i.) defining the location of object images of interest, ii) defining the boundary limits of objects, iii) characterization of object size, shape, orientation, neighbors, and surface and internal textures, and iv) application of information to artificial intelligence algorithms establishing their identity by automatic classification.*

The locating of objects in images are directly related to the image producing device and the types of preparation enhancements that have been applied. These devices utilize external excitation sources with detection of energy from emission (secondary), transmission or reflection. Alternatively, in some conditions the source is internal and alters natural or externally applied forces. In all these cases, the digital image that can be obtained is obtained as 2 or 3-dimensional 'density matrices' of pixels or voxels. In either case, surfaces of embedded objects are digitally separable from the surrounding structures because of their lower levels of detectability by the particular instruments being used, their natural density parameters and the enhancement that was applied during preparation.

Commonly, thresholds for density are used to determine the *location of objects* specified by a particular imaging technology. While density thresholds may locate a potential site, it is not usually effective in defining all the parts and limits of the structure. Other approaches are needed. Relaxation labeling can potentially solve some of these problems. For example, numerous complex structures appearing next to each other may be separated by the technique. The method can also be applied to tracking and correspondence analysis problems. However, it is highly computation-bound and requires advanced computational methods. Other algorithms will be also necessary to establish the precise limits of structures in various types of preparations.

Once the densities are located, the *boundaries are extracted* in two or three dimensions. In the case of uniformly dense objects, edge detectors and boundary search algorithms apply. The procedure in two dimensions is not trivial since threshold methods do not always apply. For example, boundaries are better described by the second derivative of density or in the case of complex objects, the boundary may have wide variations in intensity that defy all density thresholding procedures. This is common for scenes where light sources give contrasting representations of the same boundary.

Relaxation labeling is generally suited to boundary extraction since the edges can be established form data obtained from this processing. At the present time the relaxation labeling method has not been adequately tested. Supercomputers are particularly applicable to this analysis. Other approaches must also be tested and will require extensive processing time in order to locate the edges and characterize the perspective domain of objects based on information from a number of sources.

Once the boundaries of objects are located, object characterization is done by establishing quantitative parameters. These parameters define size, shape, orientation, textures, and neighbor relationships. The use of supercomputers in this task is not generally necessary but will be of advantage in development of integrated software approach bringing together all phases of image processing into automated recognition and quantitation.

2. Advanced Computation for Tomographic Approaches.

Recent developments in methods of extracting internal surfaces without physically sectioning organism now is effective for analysis of living preparations. These are ultrasound and axial tomographic approaches (CAT, PET, NMR and optical sectioning). The latter produce a solid, 3-dimensional matrix of density voxels representing the volume of structure. Improvements in resolution continue to expand the applications to finer detailed investigations. The viewing of these internal densities require digital processing with display. Hardware devices can readily define densities in sections, and even now recent developments allow objects to be selected on the basis of density and rotated in real time. Supercomputers will be necessary for further processing in order to extract additional data and differentiate between complex structures having nearly the same density.

Currently our aim for obtaining biologically useful information is limited by computing power. The amount of biologically significant data that can be extracted ultimately depends on the number of views as well as the sampling size. The ability to handle large amounts of input data (10-100 megabytes) is therefore crucial for scientific purposes. Processing steps where a supercomputer would greatly facilitate progress are: 1) combining the alignment and registration phase with the reconstruction phase on data of much higher than current capabilities, e.g. 2K x 2K; 2) investigations of iterative reconstruction algorithms which maximize functionals such as "likelihood" and "entropy"; 3) analysis of reconstructed images to extract biologically significant substructures. This step would include intercomparison of several reconstructed images which may come from different original specimens, for conserved features, and the fitting of models. For these computations, massive amounts of data would be manipulated, perhaps up to 0.5 - 1 gigabytes per reconstruction.

The projects within the tomographic area of analysis which can be considered to best benefit from the application of supercomputers are true 3-dimensional (as opposed to cross-sectional) reconstructions from projections. This study group has, in particular, identified positron emission tomography (PET), nuclear magnetic resonance (NMR) and electron microscopic tomography (EMT) as areas where research would greatly benefit from advanced computational power of the type that a supercomputing center could provide. (Optical sectioning
previously described).

a. Positron emission tomography. In advanced PET analysis, statistical information about integrated activity inside the body is collected along as many as 10^9 lines oriented in many different directions in space. From such information activity inside the body is to be estimated at, say, 10^6 points. The best methods to do this are likely to be iterative procedures (such as the "expectation-maximization technique") which optimize certain functionals (such as the "likelihood"). Research into which optimizers are most efficacious in practice and the number of iterations required are best carried out on a supercomputer.

b. NMR analysis of soft tissues. Magnetic resonance imaging is a computerized tomographic technique where existing computers are adequate for data collection. However, supercomputers can be used to advantage for post-processing of images and simulation of the physics of how the NMR signal is produced by tissue. A typical example of the data which is generated from NMR analysis of whole body regions would be a twenty slice study done in 8.5 minutes with two spin echo images from each slice. Spatial resolution is 1.7 x 1.7 x 7.0 mm with a 256 x 128 array of voxels for each echo. Up to 512 x 512 voxels could be imaged in a reasonable time. The data acquisition is currently handled using standard mini-computers with Fourier transforms done using an array processor on the mini's bus.

Magnetic relaxation rates of tissue can be measured from images and are characteristic for different normal tissues. They are also very sensitive to changes in disease but changes in specific diseases have not yet been identified. Using images of T1 and T2 new images can be calculated which are more sensitive and may represent physiological activity. As the parameters of normal and diseased tissues are cataloged, computerized tissue typing is possible.

Blood vessels can also be traced through slices, connected and rotated into more useful views. Speeded up and more sophisticated processing to reduce operator input time are needed. Images of other atomic nuclei (P^{31}, C^{13} , F^{19}) and images where the chemical spectrum for every voxel is available. These chemical images will have 10-100 times more data and thus computational aids will be needed to integrate the information from all these sources. The NMR signal has a dependence on fluid flow. Investigations of this application

are another potential for advanced computation.

c. Metabolic imaging: Reconstruction and quantitation of cellular kinetics from PET and NMR. For making observations of tissue and organ function *in vivo*, the new imaging modalities of positron emission tomography (PET and nuclear magnetic resonance (NMR) provide data sources. These must be quantitated, validated and interpreted via complex models incorporating heterogeneous flow, membrane permeation, intracellular reactions, and cascades of reaction products. An idealized data acqusition-to-analysis package for PET might include: list mode acqusition of time-of-flight data, fully 3-D reconstruction, temporal smoothing of the image sequence via the kinetic models for each of the volume events in the image.

d. Electron microscopical tomography. Studies in EM tomography could benefit from a supercomputer is the following ways. Currently, fixed specimens (up to several microns thick) are examined in a conventional transmission electron microscope (TEM & STEM) or intermediate and high voltage electron microscopes (HVEM). Magnified projections are collected in several specimen tilting intervals. The micrographs are digitized, registered and aligned. The resulting images are reconstructed in three dimensions by Fourier transforms. The results are then prepared for display and analysis.

3. Intelligent Vision Systems for 2-D and 3-D Analysis of Biological Images.

The human visual system is still many orders of magnitude ahead of the most advanced machine systems. Actually, it is quite remarkable that machine vision has advanced as far as it has in such a short period of time. This can only be due to the efforts by scientists working to extend the boundaries of computational theories of vision. Faster computers not only reduce the time required for testing of theories and algorithms on adequate data but also allow tests of processes that would not ordinarily be possible. Currently, it is not at all uncommon for image analysis experiments to require 10-20 hours of CPU time.

Experience has shown that automated industrial inspection systems can reliably perform limited sets of image analysis tasks, specifically tailored to fit a static environment. Machine recognition of the complexity of biological preparations is enhanced by providing restrictions on objects which make up the scene. These can be controlled to a large degree by the preparation procedures. Morphology is therefore a fertile testing ground for high-level computer vision development.

Evaluation of the strengths and weaknesses of the various approaches over a large number of images provides many clues for the development of theories of higher-level vision. Much work can be done in model representation and matching, multi-resolution feature extraction from complex scenes, evidential reasoning, control and search strategies, and representation of shape to name only a few. A specific task such as orientation selection requires extensive computations. Plausible models for orientation selection of biological images have been proposed that fit within the general frame work of relaxation labeling described previously. Small simulation of this model have been done on a VAX 11/780 and take on the order of 16 CPU hours per experiment. While this is prohibitive for any interactive procedure, it is feasible as a starting point to test methods that can lead to refined approaches. Without something as powerful as supercomputer facilities, experiments such as these would be impossible.

4. Summary of Graphics and Telecommunications.

Modeling, studies of molecular structure and dynamics, image processing and analysis of physiological events all require large data bases to be transferred to the computer site and

large data formats to be returned as graphical illustration. In many of these graphics applications, interactions must be made with images before the next analysis is issued at the supercomputer site. Thus, the turn around time becomes crucial, and effective utilization can only be sustained either by on-site application or by high data rate telecommunications between the laboratory and the supercomputer site. The output of analysis are images requiring medium to high resolution surface-rendered displays that generate 3-dimensional dynamic views. The display of multiple perspectives with motion create critical depth perception and object separation for investigators to conceptualize relationships and dynamic events. The combination of telecomunications and graphics capabilities in user laboratories will allow a wider range of applications in the life sciences. In order to make computer graphics an effective component of supercomputation, communication at broad band width is going to be necessary. In the immediate future, real-time interaction is not feasible unless the investigator is at the supercomputer center.

Interactive graphic operations require a minimum communication in the order of 56K baud between a group of users and the central facility. In the case of real time utilization, the rate is in excess of 1 megabit/sec for minimal data displays. In the case of real-time interactive graphics, a graphics display systems with adequate communications capability of 8 megabytes/sec is an important target for consideration for the National Science Foundation in conjunction with supercomputation. As a minimum interrum, graphics equipment at the central site is most essential for effective utilization. These would include high resolution film recorders and laser disk facilities for recording frames and shipping them to the user sites. Certainly, a real-time interactive graphics unit at the supercomputer center is manditory.

Since technology has made significant advances, the design for a telecommunication system should be a generation above the 56K level which is currently in use. Flexibility in transmission rates in order to accommodate a wide range of applications is essential. These should range from 9600 baud to over 56K baud as an immediate goal with a longer range target having a burst capability of 8 megabytes/sec by a single user. Upward of 26.5 megabytes/sec as an input-output port with direct recording onto a 200 megabyte disk is now available to user sites. This system is already interfaced to a real-time full color display system (Gould-Deanza).

Finally, there should be an interfacing of the investigators designing graphics approaches and life scientists so that many existing approaches can be implemented and additional specific ones can be designed. The direction that is taken at many junctures will depend on cross-fertilization between the two areas. Examples of applications in both image analysis and image generation are fractal methods for generators of images and deconvolutors of complexity in biological analysis. Computer graphics developers are addressing problems from an image generation point of view and many of these applications are related directly to biological problems. However before new directions can be taken, they must be implemented in software and operate in a reasonable time frame. Supercomputers can be an important bridge to developments in image processing and image generation.

REFERENCES ON COMPUTING IN THE LIFE SCIENCES

The refinement of southern bean mosiac virus in reciprocal space. A. Silva and M.G. Rossmann. *Acta Cryst. A41*: in press, 1985.

Maximum entropy and the foundations of direct methods. C. Bricogne. *Acta Cryst. A40*: 410-445, 1984.

Structure of the hamegglutin membrane glycoprotein of influenza virus at 3 Angstrom resolution. I.A. Wilson, J.J. Skehel and D.C. Wiley. *Nature 289*:368-373, 1981.

Structure of tomatoe bushy stunt virus, IV: The virus particle at 2.9 angstrom resolution. A.J. Olson, G. Bricogne and S.C. Harrison. *J. Molec. Biol. 171:* 61-93, 1983.

CHARMM: A program for macromolecular energy minimization and dynamics calculations. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan and M. Karplus. *J. Comp. Chem. 4*: 187-217, 1983.

Protein conformation, dynamics and folding by computer simulation. M. Levitt. *Ann. Rev. Biophys. Bioengin. 11:* 251-271, 1982.

Incorporation of stereochemical information into crystallographic refinement. *In: Computing in Crystallography*. Edited by R. Diamond, S. Ranaseshan and K. Venkatesan, pp. 1301-1325. Indian Academy of Sciences, 1980.

Methods and programs for direct-space exploitation of geometric redundancies. G. Bricogne. *Acta Cryst. A32*: 832-847, 1976.

Direct phase determination based on anomalous scattering. W.A. Hendrickson, J.L Smith and S. Sherify. *Methods in Enzymology 115:* in press.

Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide. W. Braun, C. Boseh, L.R. Brown N. Go and K. Wuthrich. *Biochim. Biophys. Acta. 667:* 377-396, 1981.

A distance geometry program for determining the structures of small proteins and their macromolecules for nuclear magentic resonance measurements of intramolecules 1H-1H proximities in solution. T. Harel and K. Wuthrich. *Bull. Math. Biol. 46:* 673-698, 1984.

Protein structure from nuclear magnetic resonance data. Lac repressor head piece. R. Kaptein, E. Zuiderweg, R. Scheek, R. Boelens and W. van Gunsteren. *J. Mol. Biol. 182:* in press.

Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. M. Levitt, C. Sander and P.S. Stera. *J. Mol. Biol. 181:* 423-447, 1985.

Computer simulation of the dnyamics of hydrates protein crystals and its application in protein crystallography. W.F. van Gunsteren, H.J.C. Berendsen, J. Harmans, W.G.J. Hol and J.P.M. Postma. *Proc. Natl. Acad. Sci. USA 80:* 4315-4319, 1983.

Dynamics of proteins: Elements and function. M. Karplus and J.A. McCamarran. *Ann. Rev. Biochem. 53:* 263-300, 1983.

A graphics model building and refinement system for macromolecules. T.A. Jones. *J. Appl. Cryst. 11:* 268-272, 1978.

Van der Waals surfaces in molecular modeling: Implementation with real-time computer graphics. P.A. Bash, N. Pattabiram, C. Huang, T.E. Ferrin and R. Langridge. *Science 222:* 1325-1327, 1983.

Structure of the actin molecule determined from electron micrographs of crystalline actin sheets with a tentative alignment of the molecule in the actin filament. P.R. Smith, W.E. Fowler, T.D. Pollard and U. Aebi.
J. Mol. Biol. 167: 641-660, 1983.

Three-dimensional structure of proteins determined by electron microscopy. U. Aebi, W.E. Fowler and P.R. Smith. *Ultramicroscopy 8:* 191-206, 1982.

The fluid dynamics of heart valves: Experimental, theoretical and computational methods. C.S. Peskin. *Ann. Rev. of Fluid Mechanics 14:* 235-259, 1982.

A mathematical model and numerical method for studying platelet adhesion and aggregation during blood clotting. A.L. Fogelson. *J. Comp. Physics 56:* 111-134, 1984.

The mechanical basis of morphogenesis. I. Epithelial folding and invagination. G. Odell, G. Oster, P. Alberch and B. Burnside. *Devel. Biol. 85:* 446-462, 1981.

Models of Biological Pattern Formation. H. Meinhardt. Academic Press: London, 1982.

Mechanics, morphogenesis and evolution. In: *Lectures on Mathematics in the Life Sciences*. G. Oster, G. Odell and P. Alberch. Edited by G. Oster, Vol. 13, pp. 165-255. American Mathematical Society, Providence, R.I., 1980.

A mathematically modeled cytogel cortex exhibits periodic Ca^{++}-modulated contraction cycles seen in *Physarun* shuttle streaming. G. Odell. *J. Embryol. & Exp. Morpho. 83:* Suppl. 261-287, 1984.

Some electrophysiological consequences of electrogenic sodium and potassium transport in cardiac muscle: A theoretical study. E.A. Johnson, J.B. Chapman and J.M. Kootsey. *J. Theoret. Biol. 87:* 737-756, 1980.

Electrical and biochemical properties of an enzyme model of the sodium pump. J.B. Chapman, E.A. Johnson and J.M. Kootsey. *J. Membrane Biol. 74:* 139-153, 1983.

Minimum mechanism for Na^{+}-Ca^{++} exchange: Net and unidirectional Ca^{++} fluxes as functions of ions composition and membrane potential. E.A. Johnson and J.M. Kootsey. *J. Membrane Biol.* in press, 1985

Integration and propagation of neuroelectric signals. *In: Studies in Mathematical Biology. Vol. 15 - Studies in Mathematics*. Edited by

S.A. Levin, pp. 1-66, 1976. Mathematical Association of America. Washington, D.C.

Coordination: A vector-matrix description of transformations of overcomplete CNS coordinates and a tensorial solution using the Moore-Penrose generalized inverse. A. Pellionisz. *J. Theoret. Biol. 110:* 353-375, 1984.

Tensor network theory of the metaorganization of functional geometries in the CNS. A. Pellionisz and R. Llinas. *Neuroscience, in press,* 1985.

Modeling of neurons and neuronal networks. *In: The Neurosciences: IVth Study Program.* Edited by F.O. Schmitt and F.G. Worden, pp. 525-546. MIT Press: Boston, MA, 1979.

Computer model of cerebellar Purkinje cells. A. Pellionisz and R. Llinas. *Neuroscience 2:* 37-48, 1977.

Computer model of the cerebellar cortex of the frog. A. Pellionisz, R. Llinas and D.H. Perkel. *Neuroscience 2:* 19-35, 1977.

Calcium and the mechanisms of transmitter release. S.M. Simon. Ph.D. *Dissertation.* New York University Medical Center, 1984.

Compartmentalization of the submembrane calcium activity during calcium influx and its significance in transmitter release. S.M. Simon and R.R. Llinas. *Biophys. J., in press.*

Dynamic electrotonic coupling in mammalian inferior olive as determined by simultaneous multiple Purkinje cell recording. K. Sasaki and R. Llinas. *Biophys. J. 47:* 53a, 1985.

Simultaneous sampling and analysis of the activity of multiple, closely adjacent, cerebellar Purkinje cells. J. Bower and R. Llinas. *Neurosci. Abst. 8:* 830, 1982.

Simultaneous sampling of the responses of multiple, closely adjacent, Purkinje cells responding to climbing fiber activation. J. Bower and R. Llinas. *Soc. Neurosci. Abst. 9:* 607, 1983.

Temporal patterning in simple spike discharge of Purkinje cells and its relationship to climbing fiber activity. T.J. Ebner and J.R. Bloedel. *J. Neurophysiol. 45:* 933-947, 1981.

Scene labeling by relaxation operations. Robert Hummel, A. Rosenfeld and S. Zucker. *IEEE trans. on Systems, Man and Cybernetics 6:* 420-433 1976.

On the foundations of relaxation labeling processes. R. Hummel and S. Zucker *IEEE Tr. on Pattern Analysis and Machine Intelligence 5:*267 1983.

Optical sectioning microscopy: Cellular architecture in three dimensions. D.A. Agard. *Ann. Rev. Biophys. Bioengin. 13:* 191-219, 1984.

A least-squares method for determining structure factors in three-dimensional tilted-view reconstructions. D.A. Agard. *J. Molec. Biol. 167:* 849-852, 1983.

APPENDIX

ACKNOWLEDGMENT

We wish to thank the workshop committee for helping to make the meeting a success. The plans and format were developed together with this committee consisting of Drs. Wayne Hendrickson, Stephen Harrison, Charles Gilbert, James Larimer, John Wooley, Arthur Kowalsky, Mr. Michio Chujo and Mr. Ronald Crank.

A special appreciation is extended to Digital Productions, Lucasfilms, and Phoenix Data Systems for presentations of general interest to the workshop. Contributions by Emily Nagle of Digital Productions and Robert Cook of Lucasfilms illustrated applications of advanced computation in image production for static and dynamic display. In addition, presentation of approaches for generation of images by Al Barr of Cal Tech and Arthur Olson of Scripps Institute, enhanced knowledge of how computation can be applied to problems involving graphics applications. Hardware display of graphical images in 3-dimensions by Donald Meagher of Phoenix Data gave important insights into real-time image display processing and how large arrays of data may be handled as an adjunct to supercomputation.

The group was especially appreciative of Cray Research and Control Data Corporation for sponsoring social gatherings that allowed time for meeting attendees and informal discussion of computational problems. Talks on hardware by J. Michael Flanigan of Control Data and John Aldag of Cray Research provided background on supercomputers and insight into their future direction.

We extend our appreciation to Dr. Yukihiko Karaki for presenting a perspective of supercomputation in Japan and the systems available from the Japanese computer industry. The contributions on supercomputer architecture and software approaches by Drs. J. Worlton and Simon were especially instructive in acquainting life scientists with the use of this instrumentation.

Among the outstanding presentations, there are three representing advanced computation in the life sciences that deserve special mention. A new light microscopical approach, using computation on images from optical sectioning to improve resolution greatly, was given by Dr. David Agard of University of California (San Francisco). Dr. Charles Peskin of the Courant Institute reported results of blood flow modeling in the heart. Finally, the demonstration on uses of NMR in imaging soft tissue by Dr. Lawrence Crooks of the Radiologic Imaging Lab (San Francisco) represented a potentially significant use of computation in promoting of scientific understanding throughout the biological fields.

The organizers and attendees are indebted to the National Science Foundation for the opportunity to attend the workshop and especially Drs. Larimer and Wooley for their efforts in making this meeting possible. We thank Dr. Robert Rabin for presenting the NSF perspective, and other NSF personnel and representatives for interacting in group discussions. The participation of NIH personnel and their representatives also added greatly to the success of this meeting.

Dean E. Hillman
Rodolfo Llinas
New York University Medical Center
April, 1985

WORKSHOP DISCUSSION GROUPS.

NSF Participants.

Maryanna P. Henkart, Cellular Physiology Program Director
Arthur Kowalsky, Biophysics Program Director
Pat Jost, Biophysics Program
James Larimer, Sensory Physiology & Perception Program
John C. Wooley, Biological Instrumentation Program

NIH Extramural Programs

Charles L. Coulter, Biotechnology Research Program
Jack Hahn, Biotechnology Research Program
Paul H. Lenz, Special Assistant to the Director NIH/DRR/BRTP
Susan Stimler, Division of Research Programs

Advanced Computer Architectures

Jack Hannon, Control Data Corp.
Richard Harris, Cray Research
Malvin Kalos, Courant Institute of Mathematical Sci.
Tom Toth, Control Data Corp.
Jack Worlton, Los Alamos Natl. Laboratories

Applications and Software Considerations

John Aldag, Cray Research
J. Michael Flanigan, Control Data Corp.
Leon Shiman, Whitehead Institute
Horst Simon, Boeing Computer Services

Supercomputers in Japan

Yukihiko Karaki, Univ. of Tokyo Computer Centre

X-ray Crystalography and Related Approaches

Wayne Hendrickson, Columbia University (Chairing)
Mario Amzel, Johns Hopkins School of Medicine
Gerard Bricogne, Paris
Jan Hermans, Univ. North Carolina
Robert Ladner, Genex Corp
Henry Levy, Oakridge Natl. Lab.
Arthur Olson, Scripps Clinic Res. Fdn.
Abelardo Silva, Nat. Univ. La Plata, (Argentina) & Purdue University

Magnetic Resonance in Biochemistry and Biology

Al Redfield, Brandeis University (Chairing)
Lawrence Crooks, Univ. of California, San Francisco
Irwin Kuntz, Univ. of California, San Francisco

Theoretical Analysis of Macromolecular Structure and Folding

Stephen Harrison, Harvard University (Chairing)
Walter Goad, Los Alamos Natl. Labs.
Cyrus Levinthal, Columbia University
Michael Liebman, Mount Sinai School of Medicine
Michael Levitt, Weizmann Institute
Jacob Maizel, National Institutes of Health
Harel Weinstein, Mount Sinai School of Medicine

Large Scale Recording of Dynamic Events

Rodolfo Llinas, New York Univ. Med. Ctr. (Chairing)
Robert Eisenberg, Rush Medical College
Charles Gilbert, Rockefeller University
Richard Horn, University of California (Los Angeles)
Kenneth O. Johnson, Johns Hopkins School of Medicine
Daniel Tso, Rockefeller University

Population Genetics

Newton Morton, University of Hawaii (Chairing)
Tim Bishop, University of Utah
Walter Fitch, University of Wisconsin
Robert Futrelle, University of Illinois

Mathematical Modeling

Charles Peskin, Courant Institute, NYU (Chairing)
Al Barr, California Institute of Technology
Jim Keener, University of Utah
Donald Michaels, SUNY Upstate
Gary Odell, Rensselaer Polytechnic Institue
John Rinzel, NIH
C. Frank Starmer, Duke University Medical Ctr.

Tomographic Approaches to Reconstruction

Gabor Herman, University of Pennsylvania (Chairing)
David Agard, University of California (San Francisco)
Joachim Frank, New York State Health Department
Donald Meagher, Phoenix Data Systems
Donald Olins, Oakridge National Laboratories
Michael Radermacher, New York State Department of Health

Serial Section Approaches

Don Woodward, University of Texas, Dallas (Chairing)
Charles Gilbert, Rockefeller University
Dean Hillman, New York University Medical Center

Fourier Methods in Electron Microscopy

Ueli Aebi, Johns Hopkins University
J. Frank, New York State Department of Health
Ross Smith, New York University Medical Center

2-D Image Analysis and Scenes (Computer Vision)

Bob Hummel, Courant Institute, NYU (Chairing)
Michio Chujo, New York University Medical Center
Marshall Faintich, Aerospace Center, St Louis
Shumel Peleg, University of Maryland & Isreal
Kenneth Preston, Carnegie Mellon University
Peter Selfridge, AT&T Bell Labs.
Wade Smith, University of Texas, Dallas
James Strong, NASA Goddard Space Flight Ctr.
Louis Tucker, Cornell University Medical Center
Steven Zucker, McGill University

Data Transformation for Graphics

Arthur Olson, Scripps Clinic & Research Foundation

Creation Graphics - Solids Modeling and Surface Rendering

Al Barr, California Institute of Technology
Emily Nagle, Digital Productions
Donald Meagher, Phoenix Data Systems
Robert Cook, Lucasfilms

High Speed Data Transfer (to & from SC)

Alfred Spector, Carnegie Mellon University (Chairing)
Gary Christensen, Network Systems Corp.
Dieter Fuss, Lawrence Livermore Lab.

Supercomputer Facilities

Univ. Minnesota - John Sell
Purdue University - Bill Whitson
Colorado State University - Dan Pryor
Boeing Computer Services - Horst Simon
Los Alamos Natl. Laboratories - Jack Worlton
Digital Productions - Emily Nagle
Naval Research Labs. - Judith Flippen-Anderson
Lawrence Livermore Lab. - Dieter Fuss

PROGRAM OF SYMPOSIA.

MINISYMPOSIUM- I.

Hardware and Software Considerations of Advanced Computation.
Wayne Hendrickson, Columbia University (Chairing).

NSF Initiative Promoting Use of Class VI Computers.
Robert Rabin, National Science Fdn.

Supercomputer Architecture.
Jack Worlton, Los Alamos Natl. Laboratories

Software Consideration: Supercomputer Vectorization and Optimization.
Horst Simon, Boeing Computer Services

Multi-Parallel Processor Systems
Malvin Kalos, Courant Institute of Mathematical Sciences, NYU

Supercomputers and Small Processors in an Academic Environment.
Leon Shiman, Whitehead Institute

Potential for Supercomputers in Life Sciences
John Aldag, Cray Research.
J. M. Flanigan, Control Data.

Panel Discussion- Hardware and System Software Considerations:
W. Hendrickson, Moderator
System representatives- J. Aldag, J. M. Flanigan, M. Kalos, L. Shiman, J. Worlton
User Group- A. Barr, D. Pryor, J. Sell, A. Silva, B. Witson, H. Simon

MINISYMPOSIUM- II.

Uses for Supercomputers in Image Generation
Chairperson: Arthur Olson, Scripps Clinic Res. Fdn.
Panel: Al Barr, California Inst. of Technology
Emily Nagle, Digital Productions.
Robert Cook, Lucasfilms

Alternative Graphic Technologies
Arthur Olson, Chairing
Donald Meagher, Phoenix Data Systems

Question and Panel Discussion- The Place of Supercomputers and Advanced Workstations in Image Generation and Analysis.
A. Olson, chairing- R. Cook, E. Nagle, D. Meagher, A. Smith

MINISYMPOSIUM- III.

High Speed Data Transfer to and from Supercomputers.
Chairperson: Alfred Spector, Carnegie Mellon Institute
Panel: Dieter Fuss, Lawrence Livermore Labs.
Gary Christensen, Network Systems Corp.

Questions and Panel Discussion- Communications with Supercomputer Centers.
A. Spector (chairing); G. Christensen; D. Fuss.

WORKSHOP ATTENDANCE LIST

1.	Dr. U. Aebi Department of Cell Biology & Anatomy Johns Hopkins School of Medicine 725 N. Wolfe Street Baltimore, MD 21205	301-955-8649	Determination of protein structure using EM.
2.	Dr. David Agard Department of Biochemistry University of California, San Francisco San Francisco, CA 94143	415-666-2521	3-D structural analysis of chromosomes.
3.	Mr. John Aldag Cray Research 1440 N. Northland Drive Mendota Heights, MN 55120	612-452-6650	Manager of general applications.
4.	Dr. Mario Amzel Department of Biophysics Johns Hopkins Medical School 615 Wood Basic Science Bldg. 725 N. Wolfe Street Baltimore, MD 21205	301-955-3955	3-D structure of proteins in ATP synthesis.
5.	Dr. Al Barr Computer Science Department 256-80 California Institute of Technology Pasadena, California 91125	818-356-6430	Cell motility and fluid dynamics. Image generation.
6.	Dr. James Bassingthwaighte Center for Bioengineering RF-52 University of Washington WD-12 Seattle, WA 98195	206-545-2005	Simulation resource facility in cardiovasc mass transport & exchange.
7.	Dr. David L. Beveridge Department of Chemistry Hunter College - CUNY 695 Park Avenue New York, NY 10021	212-772-5354	Monte Carlo & molecular dynamics computer simulation.
8.	Howard Bilofsky Bolt Beranck Neuman Inc. 10 Moulton St. Cambridge, MA 02238	617/497-3553	Advanced computer applications.
9.	Dr. David T. Bishop Department of Human Genetics University of Utah School of Medicine Building 531 50 N. Medical Drive Salt Lake City, UT 84132	801-581-5070	Population aggregation of common traits. Simulation of genetic traits.

10.	Gerard Bricogne LURE Batiment 209C 91405 Orsay Cedex Paris, France	33-6-941-8270	Crystallography.
11.	Mr. Gary Christensen Network Systems 7600 Boone Avenue, N. Brooklyn Park, Minnesota 55428	612-425-2202	Network system design; Communications equip. for supercomputers.
12.	Dr. Charles L. Coulter Biotechnology Research Program Branch Bldg. 31, 5B41 National Institutes of Health Bethesda, MD 20205	301-496-5411	Crystallography.
13.	Mr. Michio Chujo Department of Physiology New York University Medical Center 550 1st Avenue New york, NY 10016	212-340-5406	3-D Image analysis.
14.	Dr. John Connelly Office of Advanced Scientific Computing National Science Foundation 1800 G Street, N.W. Washington, D.C. 20550	202-357-7558	Advanced computing-NSF research support.
15.	Mr Robert Cook Lucasfilm, Ltd. Box 2009 San Rafeal, California 94912	415-499-0239	Image generation.
16.	Prof. Lawrence Crooks Radiologic Imaging Lab 400 Grandview Drive, South San Francisco, CA 94080	415-952-1369	Application of magnetic resonance to medical imaging.
18.	Dr. David J. Duchamp Physical Analytical Department Upjohn Company Kalamazoo, Michigan 49001	616-385-7766	Molecular modeling.
19.	Dr. Robert Eisenberg Department of Physiology 1750 West Harrison Rush University Chicago, Ill 60612	312-942-6467	Electrical properties of cells, tissues and channels.
20.	Dr. Marshall Faintich Defense Mapping Agency Aerospace Center 3200 South Second Street St. Louis, MO 63118-3399	314-263-4937	Computer graphics for image analysis.

21.	Dr. Walter M. Fitch Department of Physiological Chemistry, Rm. 528A Service Memorial Institute University of Wisconsin School of Medicine Madison, WI 53706	608-262-1475	Genetics.
22.	Mr. J. Michael Flanigan Marketing Consultant (HQW 09G) Control Data Corp. 8100 34th Avenue South Mailing Address/Box 0 Minneapolis, MN 55440	612-853-5641	Application of Cyber systems.
23.	Ms. Judith Flippen-Anderson Code 6030 Naval Research Laboratory Washington, D.C. 20375	202-767-2624	X-ray crystallography on biological molecules
24.	Dr. Joachim Frank Wadsworth Ctr. for Labs & Research NYS Dept. of Health Empire State Plaza Albany, NY 12201	518-474-7002	Image analysis. High voltage EM.
25.	Mr. Dieter Fuss Magnetic Fusion Energy Computer Ctr P.O. Box 5509, L 561 Lawrence Livermore Lab Livermore, CA 94550	415-422-4027	Telecommunications.
26.	Dr. Robert Futrelle Department of Genetics University of Illinois 505 S. Goodwin Avenue Urbana, Ill. 61820	217-333-4777	Mechanisms of biological shape determination.
27.	Dr. Charles Gilbert The Rockefeller University 1230 York Avenue New York, NY 10021	212-570-7670	Analysis of micro-circuitry in the cerebral cortex.
28.	Dr. Walter Goad Los Alamos National Laboratories Box 1663, Mailstop K710 Los Alamos, NM 87545	505-667-7511	Computational analysis in molecular biology & genetics.
29.	Dr. Jack Hahn Biotechnology Research Program Bldg. 31, 5B43 National Institutes of Health Bethesda, MD 20205	301-496-5411	NIH research support.

30.	Mr. Jack Hannon Regional Marketing Manager Parallel Processing Systems Control Data Corp. 1900 Market Street, 5th Floor Philadelphia, PA 19103	215-854-1020	Representative of Cyber systems.
31.	Mr. Richard Harris Eastern Sales Representative Cray Research 11710 Beltsville Drive Beltsville MD. 20705	301-595-5100	Cray Representative.
32.	Dr. Stephen Harrison Biophysics Department 112 Fairchild Bldg. Harvard University 7 Divinity Avenue Cambridge, MA 02138	617-495-4090	Crystallographic studies of macromolecular structures.
33.	Dr. Wayne Hendrickson Department of Biochem & Molecular Biophysics Columbia University New York, NY 10032	212-694-3456	Protein crystallography.
34.	Dr. M. P. Henkart Cellular Physiology Program Director Division of Cellular Biosciences National Science Foundation 1800 G. Street, Rm. 325 Washington, D.C. 20550	202-357-7377	NSF research support.
35.	Dr. Gabor T. Herman Department of Radiology Medical Imaging Section 3400 Spruce Street University of Pennsylvania Philadelphia, PA 19104	215-662-6784	Medical image processing.
36.	Dr. Jan Hermans Department of Biochemistry University of North Carolina Chapel Hill, NC 27514	919-966-4644	Hydration of proteins. Molecular dynamics and thermodynamic functions. Proteins in blood coagulation.
37.	Dr. Dean E. Hillman Department of Physiology & Biophysics New York University Medical Center 550 First Avenue New York, NY 10016	212-340-5417	Quantitation of spatial relationship in CNS circuitry & substructure.

38.	Prof. Richard Horn Department of Physiology UCLA School of Medicine Los Angeles, CA 90024	213-825-5556	Mechanisms of the gating of ionic channels.
39.	Dr. Robert Hummel New York University Courant Institute of Mathematical Sciences 251 Mercer Street New York, NY 10012	212-460-7282	Computer vision.
40.	Dr. Kenneth O. Johnson Department of Neuroscience Johns Hopkins School of Medicine 725 North Wolfe Street Baltimore, MD 21205	301-955-2730	Neurophysiology.
41.	Dr. Pat Jost Staff Associate Biophysics Program Molecular Biosciences, Rm 325 National Science Foundation Washington D.C., 20550	202-357-7777	NSF research support.
42.	Dr. Malvin Kalos New York University Courant Institute of Math Sciences 251 Mercer Street New York, NY 10012	212-460-7480	Statistical physics. Quantum physics. Computer simulation.
43.	Dr. Yukihiko Karaki Computer Centre University of Tokyo Yayoi 2-11-16, Bunkyo Tokyo, JAPAN 113	81-3-8122111,	Supercomputer applications.
44.	Dr. Joyce Kaufman Department of Chemistry Johns Hopkins University Baltimore, MD 21218	301-338-7468	Drug-receptor interactions. Spatial interactions of atoms.
45.	Dr. James Keener 205 Math Building University of Utah Salt Lake City, Utah 84112	801-581-6089	Wave propagation in excitable media.
46.	Dr. J. Mailen Kootsey Director, National Biomedical Resource Duke University Medical Center Box 3709 Durham NC 27710	919-681-3048	Simulation Modelling

47.	Dr. Arthur Kowalsky Biophysics Program Director Division of Molecular Biosciences National Science Foundation 1800 G Street, Rm. 325 Washington, D.C. 20550	202-357-7777	NSF research support.
48.	Dr. Irwin Kuntz Department of Pharmaceutical Chemistry University of California San Francisco, CA 94143	415-666-1937	Determination of molecular structure using NMR.
49.	Dr. James Larimer Sensory Physiology & Perception, Rm. 320 National Science Foundation 1800 G. Street, N.W. Washington, D.C. 20550	202-357-7428	NSF research support.
50.	Dr. Robert Ladner Genex Corp 16020 Industrial Drive Gaithersburg, MD 20877	301-258-0552 x339	X-ray crystallography. Macromolecular modelling.
51.	Dr. Paul H. Lenz Special Assistant to the Director NIH/DRR/BRTP 9000 Rockville Pike Bldg. 31, Rm 5b43 Bethesda, Maryland 20205	301-496-4235	NIH research resources.
52.	Prof. Michael Levitt Department of Chemical Physics Weizmann Institute of Science Rehovot, Israel	011-972-8-482365	Protein conformation and dynamics. Computer graphics.
53.	Dr. Cyrus Levinthal Department of Biology Columbia University Fairchild Bldg. New York, NY 10025	212-280-2439	Advanced computer design.
54.	Dr. Allan H. Levy Prof & Head, Dept. Medical Info. Science. University of Illinois College of Medicine at Urbana - Champaign 1408 West University Avenue Urbana, Ill. 61801	217-333-9181	Health Info. systems. Medical data base management.
55.	Dr. Henry Levy University of Tennessee P.O. Box Y Oakridge Grad School of Biomed Sciences Oakridge National Labs Oakridge, Tennessee 37830	615-574-1265	Crystallography.

56.	Dr. Michael N. Liebman Department of Pharmacology Mt. Sinai School of Medicine 1 Gustave Levy Place New York, NY 10029	212-650-7018	Molecular modeling of proteins. Drug receptor binding.
57.	Dr. Rodolfo Llinas Department of Physiology & Biophysics New York University Medical Center 550 First Avenue New York, NY 10016	212-340-5415	Analysis of multiple electrode recordings in CNS.
58.	Dr. Jacob Maizel Bldg. 6, Rm. B2-27 National Institutes of Health Bethesda, MD 20205	301-496-4681	Genetic sequence for protein synthesis.
59.	Dr. Donald Meagher Pheonix Data System 80 Wolf Road Albany, NY 12205	518-459-6202	3-D Computer graphics. Medical imaging. Solids modelling.
60.	Dr. Donald C. Michaels Department of Pharmacology SUNY Upstate 766 Irving Avenue Syracuse NY 13210	315 473-5144	Math modelling. Cardiac electrophysiol.
61.	Dr. Newton Morton Populations Genetics Lab University of Hawaii 1980 Eastwest Road Honolulu, Hawaii 96822	808-948-7186	Origin of chromosomal abnormalities.
62.	Emily K. Nagle Digital Productions 3416 South Los Cieinega Los Angeles, CA 90016	213-938-1111	Image generation.
63.	Dr. Cynthia Null Executive Director Federation of Behavioral Psychological and Cognitive Sciences 1200 Seventeenth St. NW Washington, D.C. 20036	202-955-7758	Congnitive Sciences
64.	Dr. Gary Odell Amos Eaton Building, Rm. 328 Math Science Department Rensselaer Polytechnic Institute Troy, NY 12181	518-266-6899	Math modelling. Computer graphics.

65.	Dr. Donald Olins University of Tennessee P.O. Box Y Oakridge Grad School of Biomed Sci. Biology Division Oakridge National Labs Oakridge, Tennessee 37830	615-574-1265	Structural cell biology. Eukaryotic chromosome structure & function.
66.	Dr. Arthur Olson Scripps Clinic and Research Foundation Department of Molecular Biology, MB-3 10666 North Torrey Pines Road La Jolla, CA 92037	619-457-9702	Development & application of computer graphic approaches.
67.	Dr. Helmuth F. Orthner Office of Academic Computer Science George Washington University Medical Center 2300 K Street, N.W. Washington, D.C. 20037	202-676-2692	Information Processing.
68.	Dr. Shumel Peleg Center for Automation University of Maryland College Park, MD 20742	301-454-4526	Image analysis.
69.	Dr. Charles Peskin New York University Courant Institute of Math Sciences 251 Mercer Street New York, NY 10012	212-460-7161	Fluid dynamics. Membrane properties.
70.	Dr. Kenneth Preston Dept. Electrical Eng. Carnegie Mellon University Pittsburgh, PA 15213	412-578-2462	Medical image analysis.
71.	Dr. Dan Pryor Institute for Computational Studies Colorado State University P.O. Box 1852 Fort Collins, CO 80522	303-491-6659	Capabilities of supercomputing environment.
72.	Dr. Robert Rabin Biological Directorate of NSF National Science Foundation 1800 G. Street, N.W. Washington, D.C. 20550	202-397-9894	NSF research support.
73.	Dr. Michael Radermacher Wadsworth Ctr. for Labs & Research New York State Department of Health Albany, NY 12201	518-474-5821	Image analysis. High voltage EM.

74.	Dr. Al Redfield Department of Biochemistry Brandeis University Waltham, MA 02254	617-647-2713	NMR studies of macromolecules with emphasis on RNA.
75.	Dr. John Rinzel National Institutes of Health Bldg. 31, Rm. 4B54 Bethesda, MD 20205	301-496-4325	Math modelling in neurobiology.
76.	Dr. Peter G. Selfridge AT&T Bell Labs Rm 4F 625 Homedale, NJ. 07733	201-949-2521	Image analysis.
77.	Mr. John Sell Lauderdale Computer Facility 2520 Broadway Drive Lauderdale, MN 55113	612-373-7878	Manager supercomputer center.
78.	Dr. Leon Shiman Whitehead Institute Nine Cambridge Center Cambridge, MA 02142	617-258-5137	Structure of genes. Molecular modelling.
79.	Dr. Abelardo Silva Departamento de Fisica Facultad de Ciencias Exactas Universidad National de La Plata C.C. 67, 1900 La Plata, ARGENTINA	54-21-39061	X-ray crystallography.
80.	Dr. Horst Simon Boeing Computer Services, MS 9C-01 565 Andover Park West Seattle, WA 98188	206-575-5439	Structural analysis. Aerodynamics and petroleum reservoir.
81.	Dr. P. Ross Smith Department of Cell Biology New York Univeristy Medical Center 550 1st Avenue New York, NY 10016	212-340-5356	Image analysis of protein EM.
82.	Dr. Wade Smith Department of Cell Biology Univ. of Texas Health Science Center at Dallas 5323 Harry Hines Blvd. Dallas, Texas 75235	214-688-2483	Image analysis.
83.	Dr. Alfred Spector Department of Computer Science Carnegie Mellon University Pittsburgh, PA 15213	412-578-2583	Telecommunication.

84.	Dr. C. Frank Starmer, Jr., Professor of Computer Science Duke University Medical Center Box 3181 Durham, NC 27710	919-684-6804	Cardiac electrophysiol.
85.	Dr. Sue Stimler Biotechnology Research Program Branch Building 31, Rm 5B41 National Institutes of Health Bethesda, MD. 20205	301-496-5411	NIH research resources.
86.	Dr. James Strong NASA Goddard Space Flight Ctr. Code 65 Greenbelt MD 20771	301-344-7000 x 9535	Image analysis.
87.	Tom Toth (HQS 09S) Market Manager, Cyber 200 series Control Data Corp. 8100 34th Avenue South Mailing Address/ Box O Minneapolis, MN 55440	612-853-5641	Representative of Cyber systems.
88.	Daniel Tso The Rockefeller University 1230 York Avenue New York, NY 10021	212-570-7671	Image analysis. Multiple electrode recording.
89.	Dr. Louis Tucker Department of Neurobiology Cornell Medical School 411 E. 69th Street New York, NY 10021	212-472-5594	Computer vision and artificial intelligence.
90.	Dr. Harel Weinstein Department of Pharmacology Mt. Sinai School of Medicine 1 Gustave Levy Place New York, NY 10029	212-650-7018	Relation between molecular structure of drugs and neurotransmitters.
91.	Mr. Bill Whitson Computing Center Mathematical Science Bldg., Rm. B78 Purdue University W. Lafayette, IN 47907	317 494-1787	Supercomputer User training.
92.	Dr. Donald Woodward Department of Cell Biology Univ. of Texas Health Science Ctr 5323 Harry Hines Blvd. Dallas, Texas 75235	214-688-2483	Neuroanatomy. Neurophysiology. Image analysis and reconstruction.

93. Dr. John C. Wooley 202-357-7652
Division of Physiology, Cellular & Molecular Biology, Rm. 325
National Science Foundation
1800 G. Street, N,W.
Washington, D.C. 20550
NSF research support.

94. Dr. Jack Worlton 505-667-1449
Los Alamos National Laboratories
Box 1663, Mailstop B260
Los Alamos, NM 87545
Supercomputer center.

95. Dr. William S. Yamamoto 202-676-3871
Department of Clinical Medicine
George Washington University Medical Center
Washington, D.C. 20037
Mathematical computing.

96. Dr. S.W. Zucker 514-392-5412
Department of Electrical Engineering
McGill University
3480 University Street West
Montreal, Quebec CANADA H3A 2A7
Image analysis.

1. Computer Architecture

HIGH SPEED IMAGE PROCESSORS

S.R. John
General Dynamics
August 1985

SYSTEM DEFINITION AND DESIGN OVERVIEW

INTRODUCTION

- **SYSTEM DEFINITION**
  - PROVIDE HARDWARE AND SOFTWARE TO PERFORM STEREO MENSURATION FOR:
    - GENERATION OF AEROTRIANGULATION AND TARGET DATA
    - STEREO COMPILATION OF TERRAIN ELEVATION, FEATURE AND AIRFIELD INFORMATION
  - PROCESSING TO BE IN A DIGITAL ENVIRONMENT
  - PERFORM THE FUNCTIONS OF BOTH:
    - STEREO COMPARATOR
    - ANALYTIC STEREOPLOTTER

T065T-010

OBJECTIVES

- COLLECT DIGITAL TERRAIN ELEVATION DATA
- COLLECT DIGITAL TARGETING INFORMATION
- COLLECT DIGITAL AIRFIELD INFORMATION

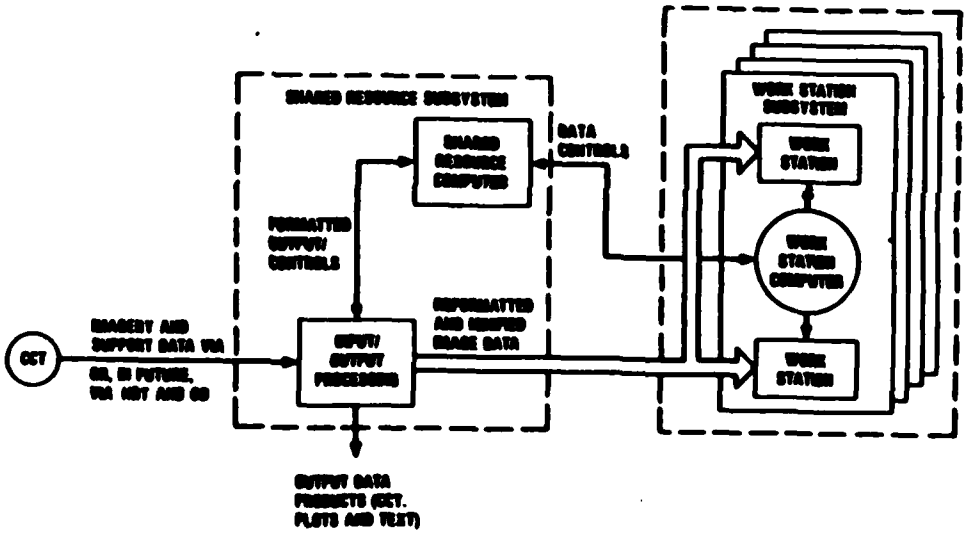
T065T-011

OUTPUT PRODUCTS

1.1.4

- **PRIMARY OUTPUT PRODUCTS**
  - DIGITAL TERRAIN ELEVATION DATA (DTED)
  - TARGET AREA
  - AIRFIELD INFORMATION DATA (AAFIF)
- **SECONDARY OUTPUT PRODUCTS**
  - IMAGE PATCHES AND ASSOCIATED SUPPORT DATA
  - PRINTED REPORTS
  - PLOTTED DATA

SYSTEM DIAGRAM



KEY PERFORMANCE REQUIREMENTS

1.1.5

- SYSTEM CAPACITY
- IMAGE PROCESSING
- IMAGE DISPLAY
- SYSTEM SPEED
- MODES OF OPERATION
- POSITIONING AND MENSURATION ACCURACY
- PHOTOGRAMMETRIC OPERATIONS

T2057-015

SUMMARY OF KEY REQUIREMENTS

- SYSTEM CAPACITY
  - TWO 4.0 X 10^8 PIXEL IMAGES PLUS 50 1024 X 1024 PATCHES OR EQUIVALENT
  - REQUIRES 86% OF IMAGE DISK LEAVING 5% FOR BAD TRACKS AND 9% SPARE
- IMAGE PROCESSING
  - FILTER - 7 X 7 CONVOLUTION
  - TONAL ENHANCEMENT - BIAS GAIN + TTC
  - ROTATION - CONTINUOUS OVER 360 DEGREES AND 90-DEGREE STEPS
  - CURSOR - BOTH FIXED AND MOVING CURSOR MODES
  - GRAPHS - LOCKED TO IMAGE
  - MINIFICATION - 1, 2, 4, 8, 16 AND 32 TO 1 REDUCTIONS
  - MAGNIFICATION - VARIABLE FROM 0.8:1 TO 5:1

T2057-010A

SUMMARY OF KEY REQUIREMENTS (CONTINUED)

1.1.6

- **IMAGE DISPLAY**
  - STEREO - 512 X 512 PIXELS VIA CUSTOM-DESIGNED VIEWER
  - OVERVIEW - 1024 X 1024 PIXELS
- **SYSTEM SPEED**
  - ROAM - > 200 PIXELS / SEC WHILE MAINTAINING STEREO
  - HIGH SPEED SLEW - MOVE TO ANY POINT IN < 2 SEC
  - AUTO DTED COLLECTION - ≥ 200 POINTS / SEC (PREDICT ~ 330 / SEC)

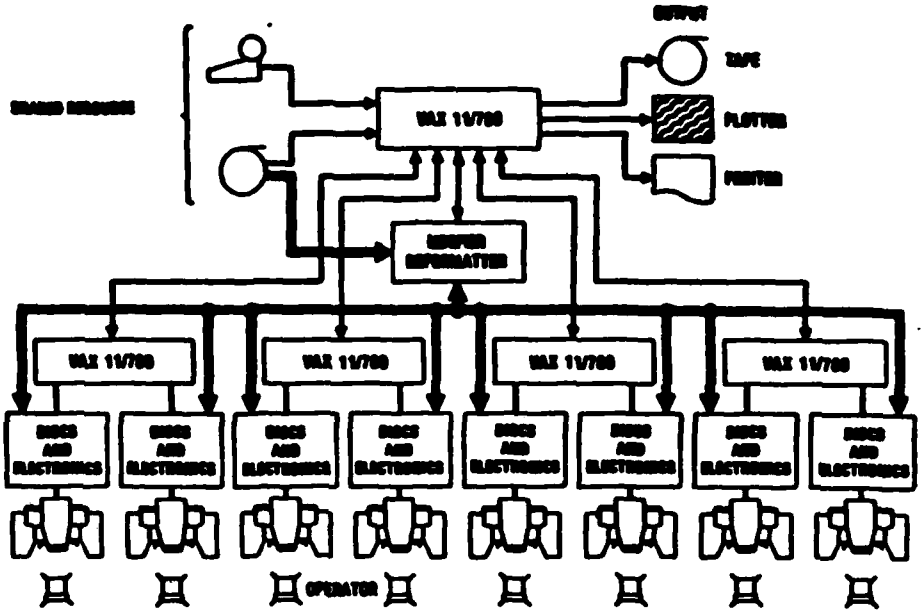
7055Y-017

SUMMARY OF KEY REQUIREMENTS (CONTINUED)

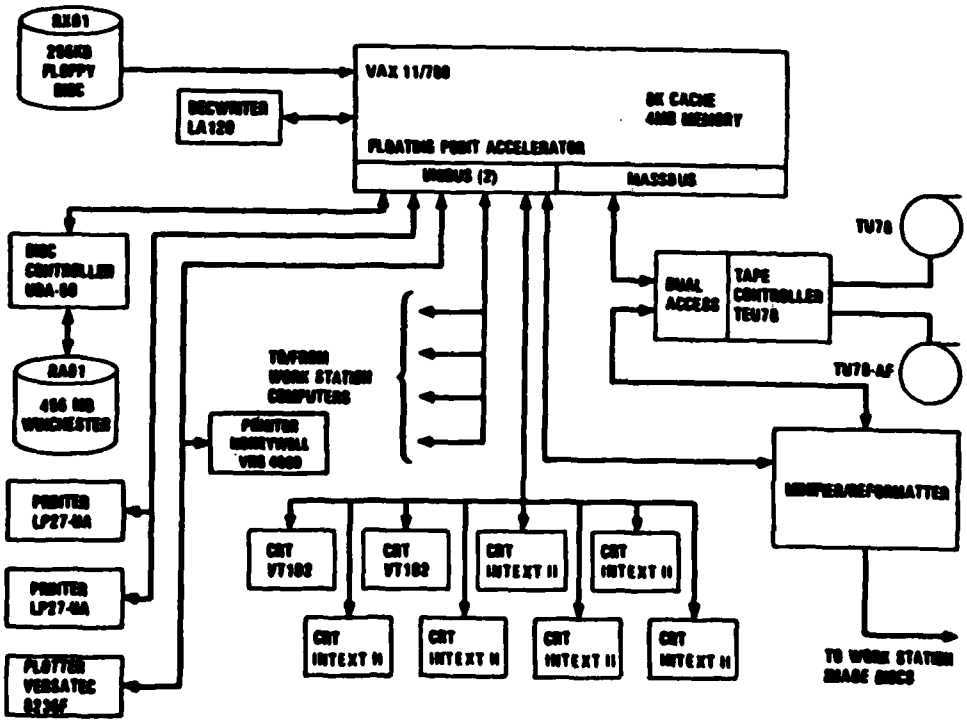
- **MODES OF OPERATION**
  - MANUAL POSITIONING
  - SEMI-AUTOMATED STEREO POSITIONING
  - AUTOMATIC MONOSCOPIC POSITIONING
  - AUTOMATED STEREO POSITIONING
  - AUTOMATED TIE / DIAGNOSTIC POINT POSITIONING
- **POINT MEASUREMENT (CONTROL, FIDUCIAL, RESEAU)**
  - ORIENTATION (INTERIOR, EXTERIOR)
  - COLLECTION (TARGET, AAFIF, DTED, TIE AND DIAGNOSTIC)
  - EDITING (ALL OF ABOVE)

7055Y 017 1A

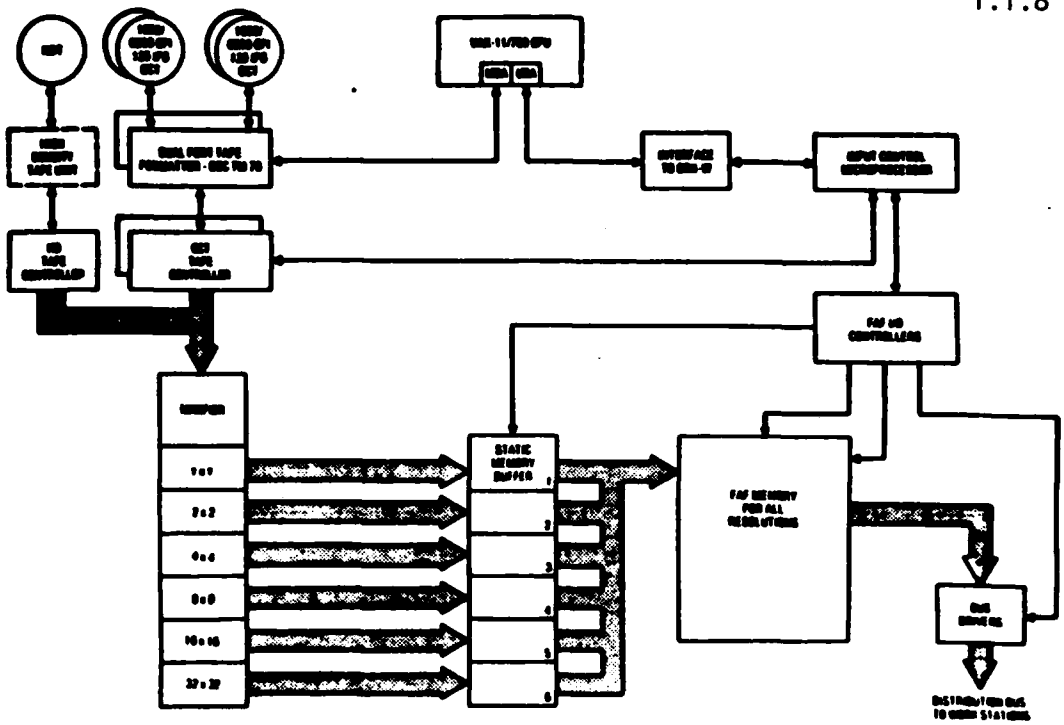
SYSTEM ARCHITECTURE



SHARED RESOURCE SUBSYSTEM



MINIFIER / REFORMATTER BLOCK DIAGRAM



WORK STATION SUBSYSTEM

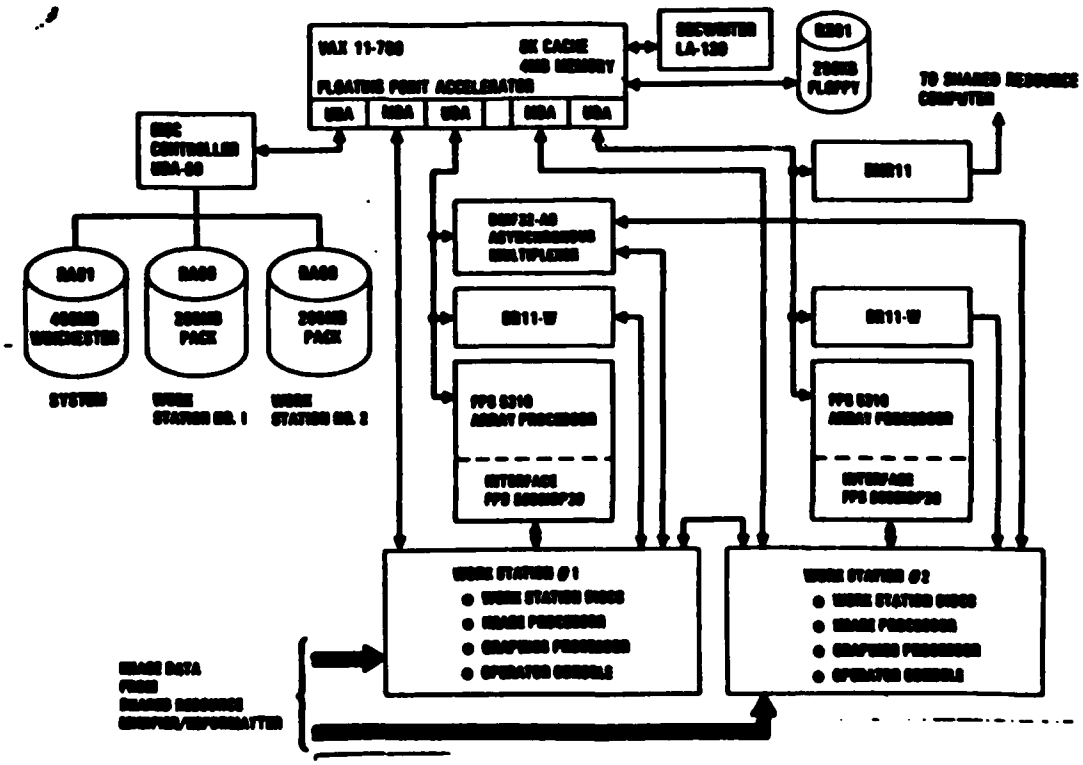
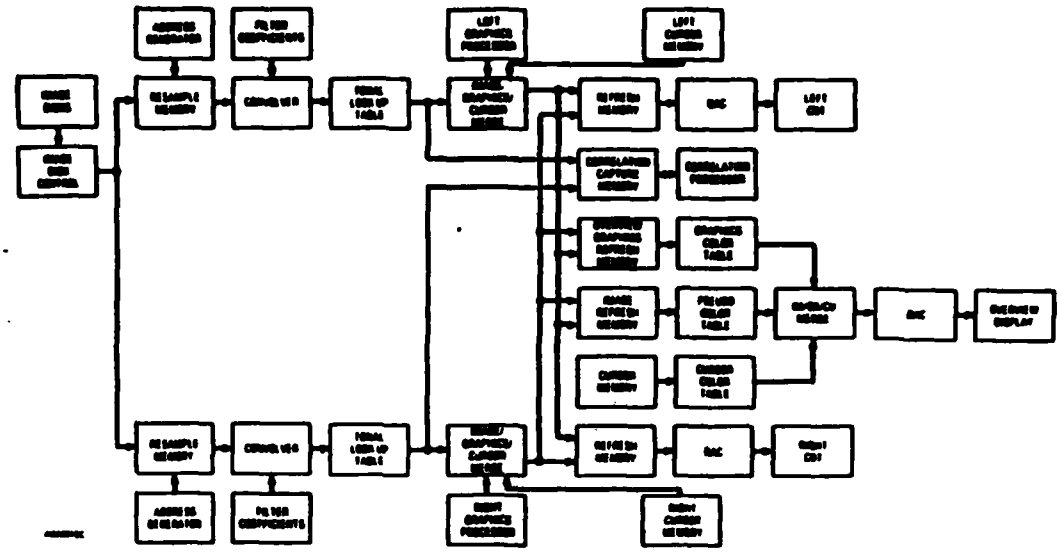
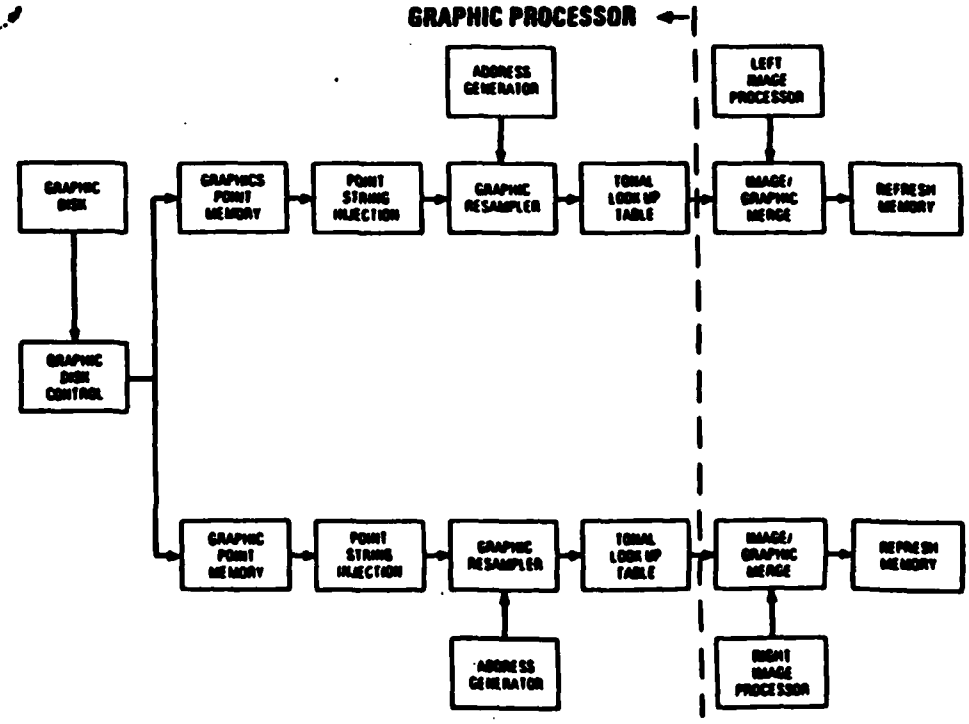


IMAGE AND DISPLAY PROCESSING BLOCK DIAGRAM



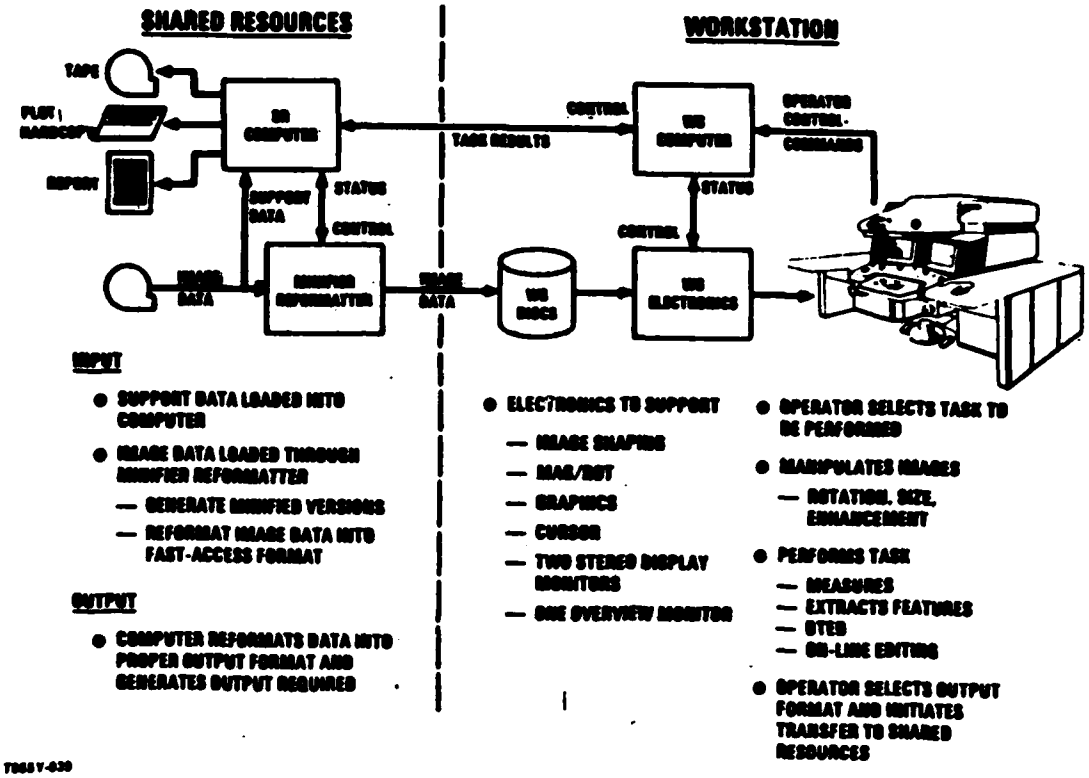
7055Y-037

GRAPHIC PROCESSING BLOCK DIAGRAM



7055Y-038

OPERATIONAL FLOW



T9661Y-030

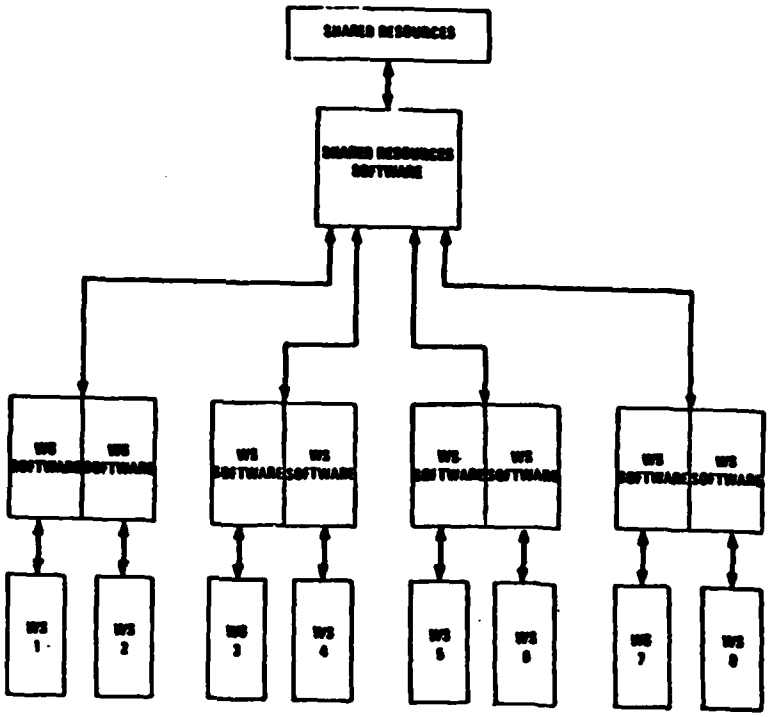
DESIGN PROCESS ENHANCED BY WORKING GROUPS

- ***GOVERNMENT-GDE-HAI TEAMS FORMED TO ADDRESS SPECIFIC AREAS***
  - ***CONSOLE WORKING GROUP***
  - ***OPERATIONS WORKING GROUP***
  - ***TEST PLAN WORKING GROUP***
  - ***OPERATOR PHYSICAL INTERFACE, CONSOLE CONFIGURATION, AND LAYOUT***
  - ***OPERATIONAL INTERFACE WITH SYSTEM***
  - ***SYSTEM TEST PLAN AND PAT, FAT PROCEDURES***
- ***THE TEST PLAN WORKING GROUP WILL CONTINUE AFTER CDR***

T9661Y-040

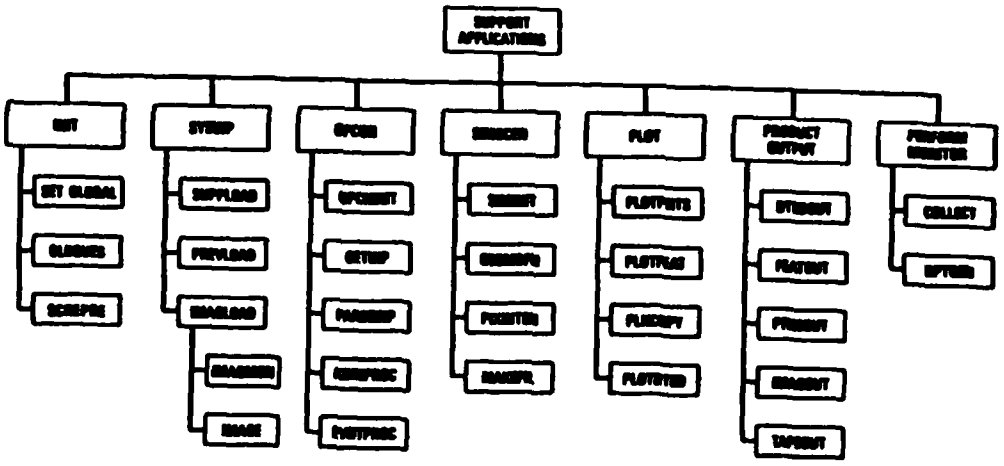
APPLICATION SOFTWARE ALLOCATION

1.1.11



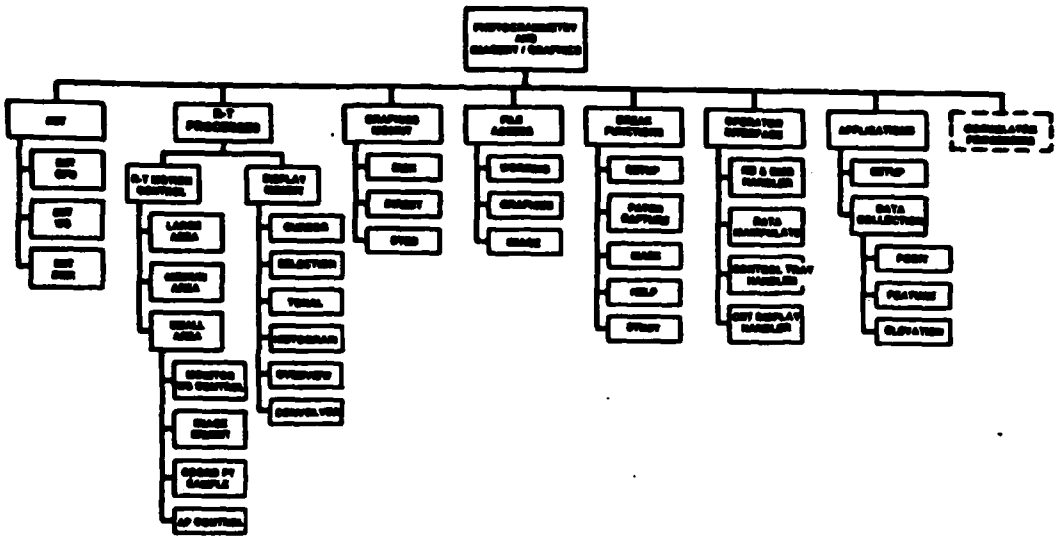
T055Y-042

SHARED RESOURCE SOFTWARE HIERARCHY

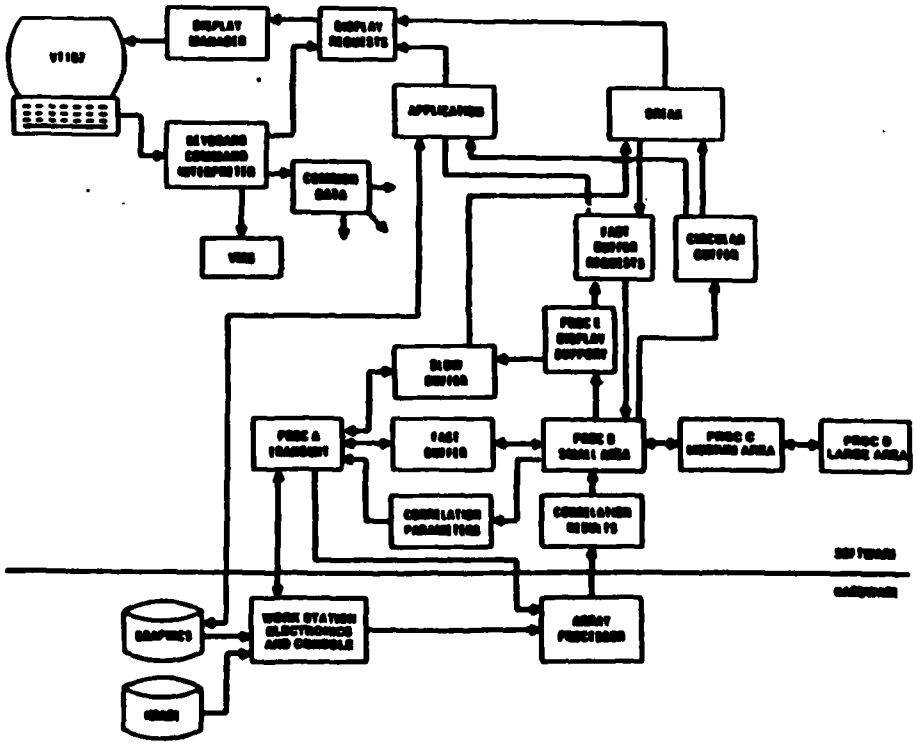


T055Y-043

WORK STATION SOFTWARE HIERARCHY



WORK STATION SOFTWARE BLOCK DIAGRAM

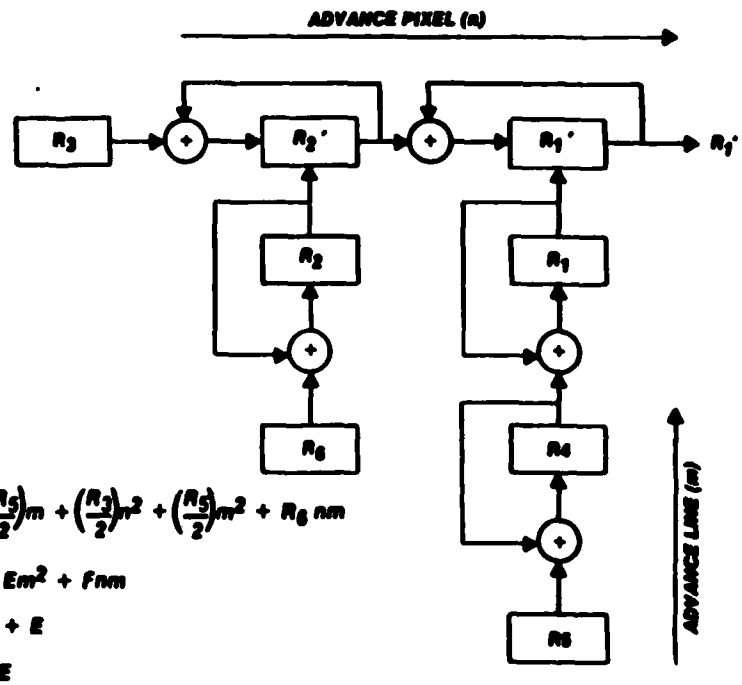


PIPELINE VS CONVENTIONAL PROCESSING

- **CONVENTIONAL**
  - PERFORMS ONE COMPUTATION AT A TIME
  - USES COMMERCIAL GENERAL PURPOSE EQUIPMENT
- **PIPELINE**
  - PERFORMS MANY COMPUTATIONS AT SAME TIME
  - USES SPECIAL PURPOSE HARDWARE

T0867Y-047

ADDRESS GENERATOR CONCEPT



$$R_1' = R_1 + \left(R_2 - \frac{R_3}{2}\right)n + \left(R_4 - \frac{R_5}{2}\right)m + \left(\frac{R_3}{2}\right)n^2 + \left(\frac{R_5}{2}\right)m^2 + R_6\, nm$$

$$R_1' = A + Bn + Cm + Dn^2 + Em^2 + Fnm$$

$R_1 = A$ $\qquad$ $R_4 + C + E$

$R_2 = B + D$ $\qquad$ $R_5 = 2E$

$R_3 = 2D$ $\qquad$ $R_6 = F$

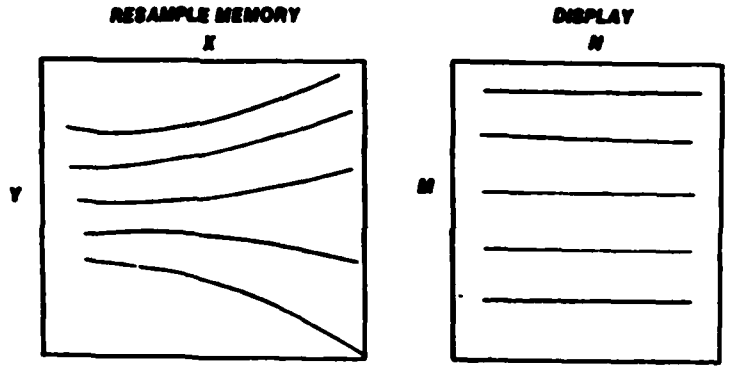
T0867Y-048

TWO GENERATORS CREATE RESAMPLED IMAGE

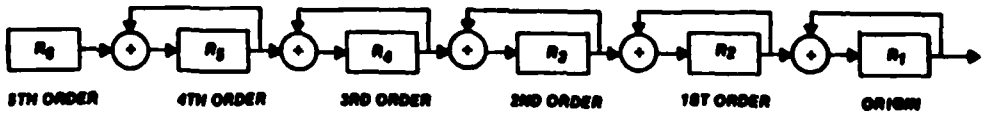
1.1.14

$$X = A_x + B_x n + C_x m + D_x n^2 + E_x M^2 + F_x nm$$

$$Y = A_y + B_y n + C_y m + D_y n^2 + E_y m^2 + F_y nm$$



ADDTIONAL TERMS CAN BE ADDED TO GENERATE HIGHER ORDER RESHAPING



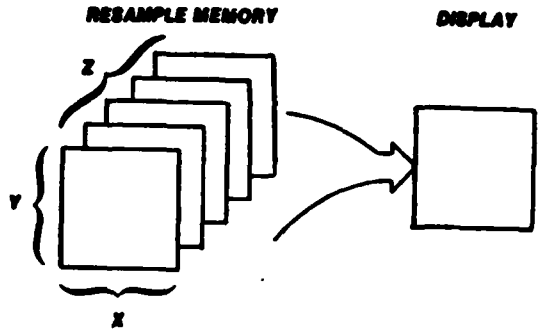
THREE GENERATORS CAN BE USED TO DISPLAY FROM A 3D RESAMPLE MEMORY

1.1.15

$X = A_x + B_x n \cdots\cdots$

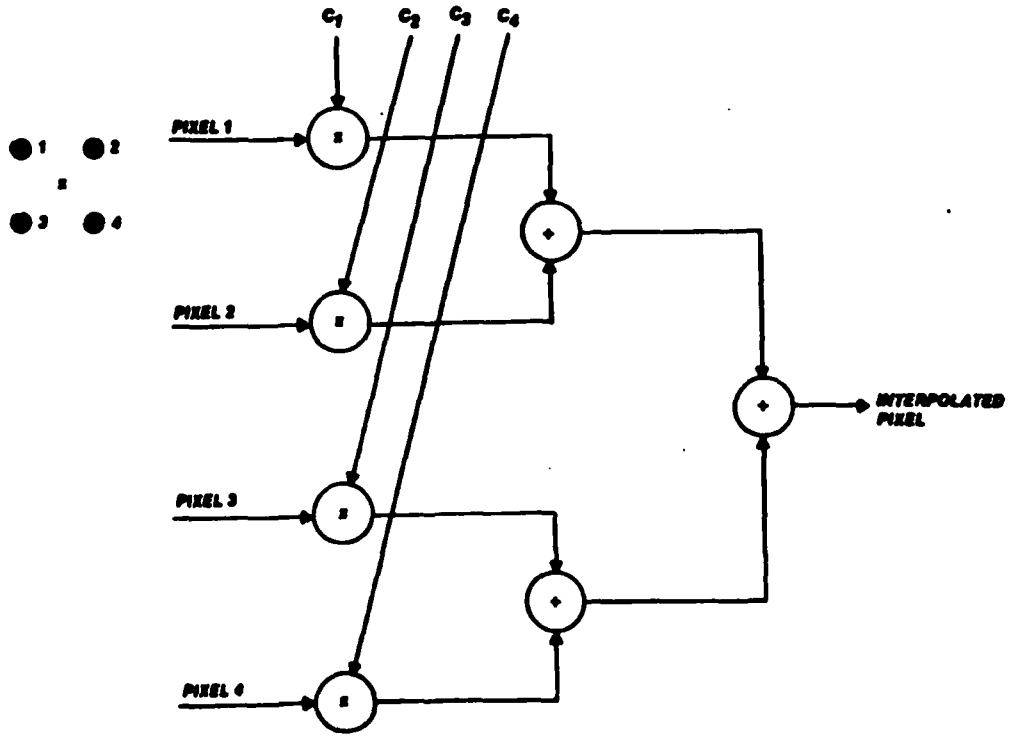
$Y = A_y + B_y n \cdots\cdots$

$Z = A_z + B_z n \cdots\cdots$



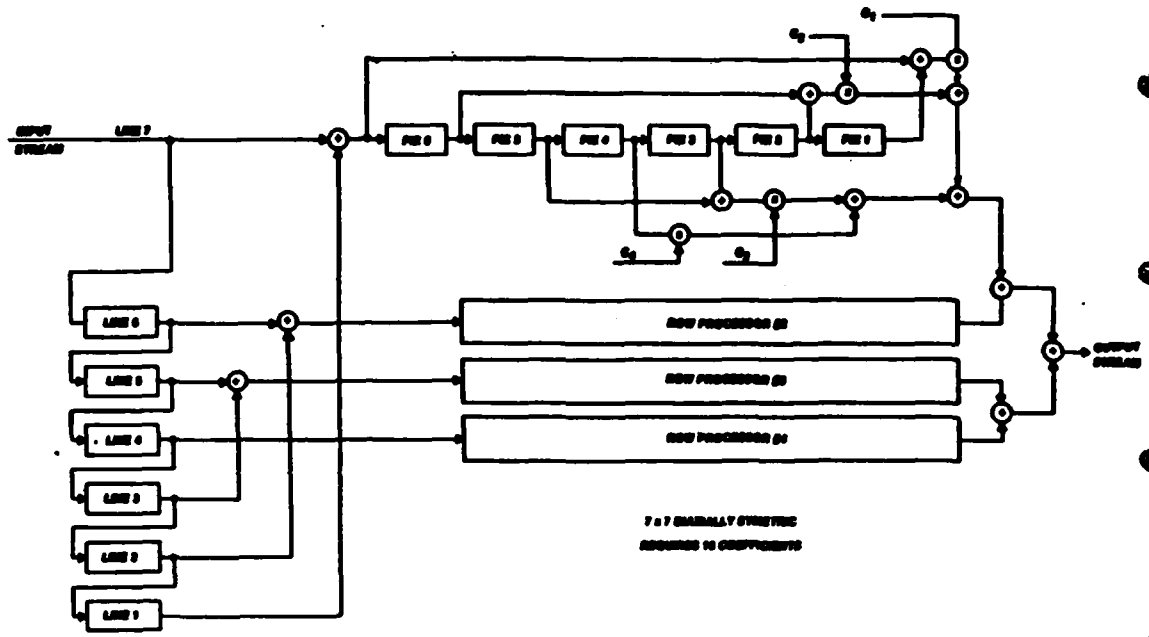
T8567-051

INTERPOLATOR 2 x 2 BILINEAR



T8567-052

CONVOLVER



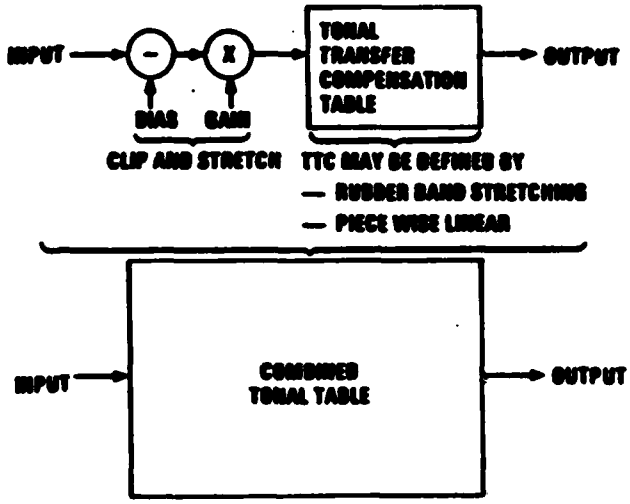
T85SY-053

CONTRAST ENHANCEMENT

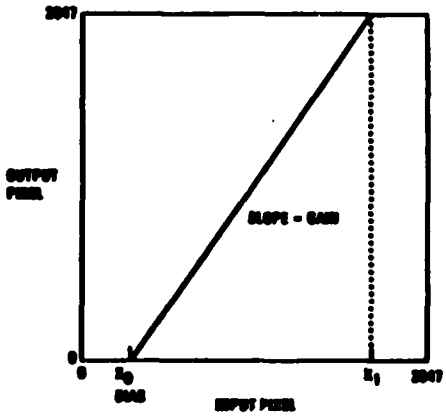
- CLIP AND STRETCH
- HISTOGRAM BASED
- RUBBER BAND STRETCHING
- PIECE WISE LINEAR

T85SY-054

GENERAL FORM



CLIP AND STRETCH



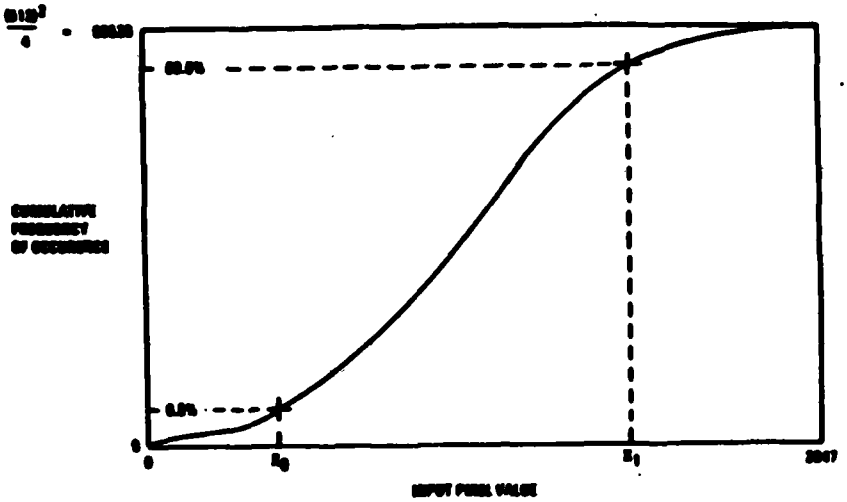
CLIP SECTION X_0 TO X_1 AND STRETCH OVER FULL OUTPUT RANGE

$$\text{BIAS} = X_0$$

$$\text{GAIN} = \frac{2047}{X_1 - X_0}$$

$$\text{OUTPUT} = \left(\frac{2047}{X_1 - X_0}\right) \cdot \text{INPUT} - X_0$$

HISTOGRAM BASED

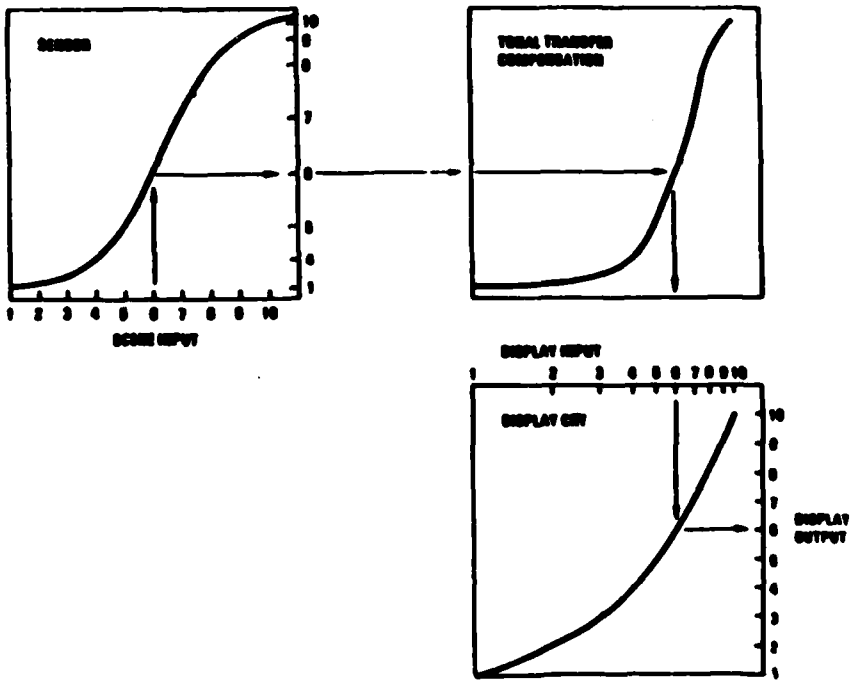


USE X_0 AND X_1 TO DETERMINE BIAS AND GAIN FOR CLIP AND STRETCH

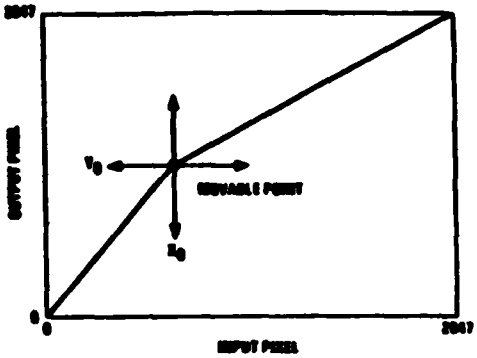
$$\text{BIAS} = X_0$$

$$\text{GAIN} = \frac{2047}{X_1 - X_0}$$

TONAL COMPENSATION CONCEPT



RUBBER BAND STRETCH



FOR INPUT $\leq X_0$

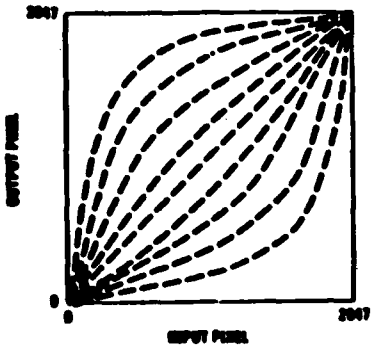
$$\text{OUTPUT} = \left(\frac{Y_0}{X_0}\right) * \text{INPUT}$$

FOR INPUT $> X_0$

$$\text{OUTPUT} = \left(\frac{2048 - Y_0}{2048 - X_0}\right) * \text{INPUT}$$

NOTE: MAY BE PRECEDED BY A CLIP AND STRETCH

PIECE-WISE LINEAR



A FAMILY OF TONAL TRANSFER COMPENSATION CURVES

EACH CURVE IS DEFINED BY 128 BREAKPOINTS

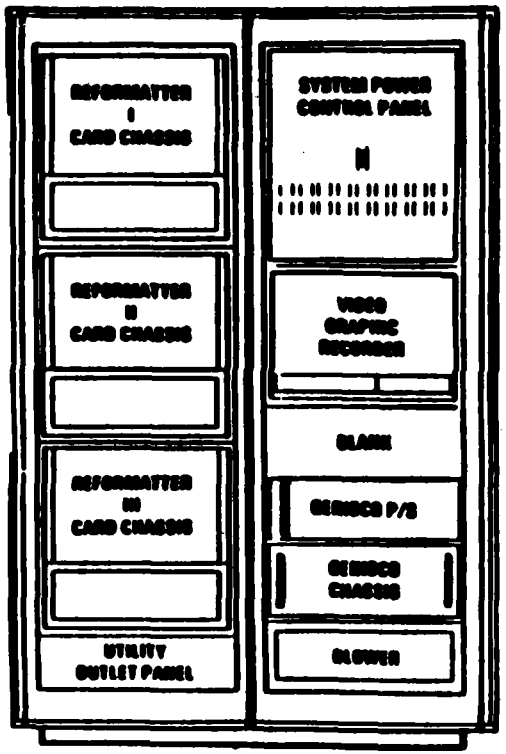
NOTE: MAY BE PRECEDED BY A CLIP AND STRETCH

1.1.20

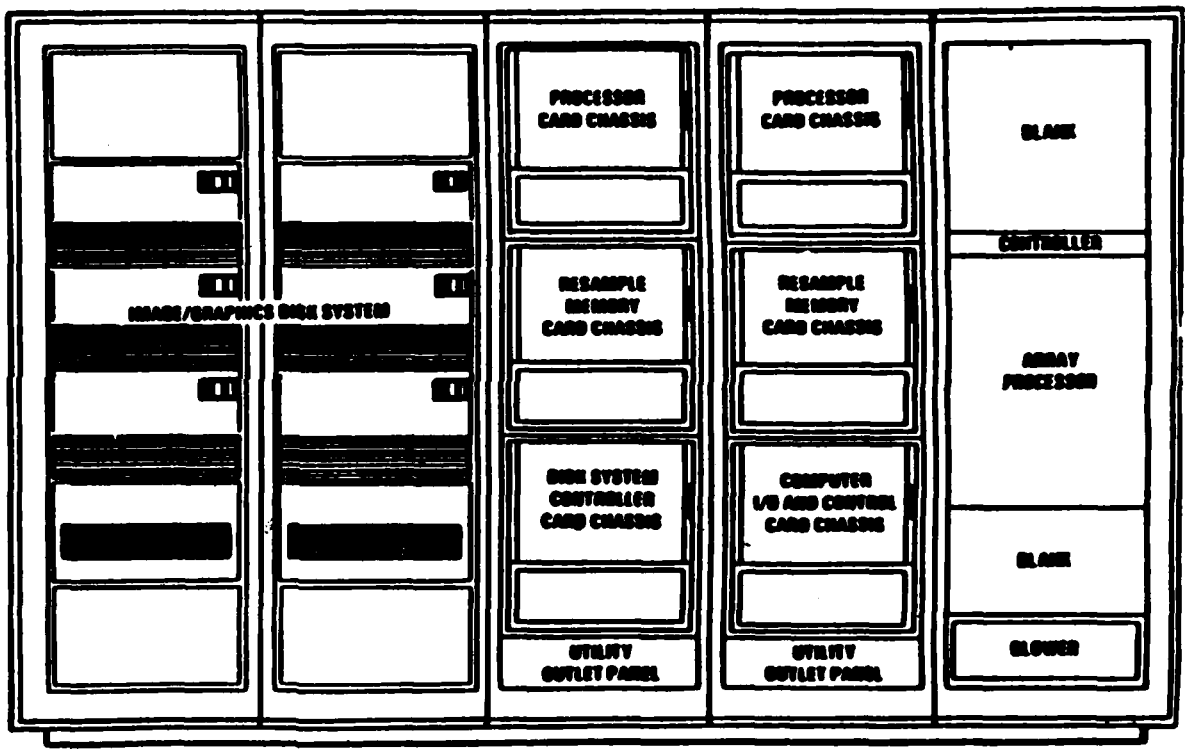
HARDWARE

ELECTRONICS

SHARED RESOURCES



WORK STATION



SR - 23 BOARDS

WS - 56 BOARDS
8WS - 448 BOARDS

TOTAL SYSTEM: 471 BOARDS
36 UNIQUE DESIGNS

T855V-062

CLUSTER IMAGE SUBSYSTEM

- HIGH STORAGE CAPACITY
- HIGH TRANSFER RATE
- HIGH RELIABILITY
- LOWER COST

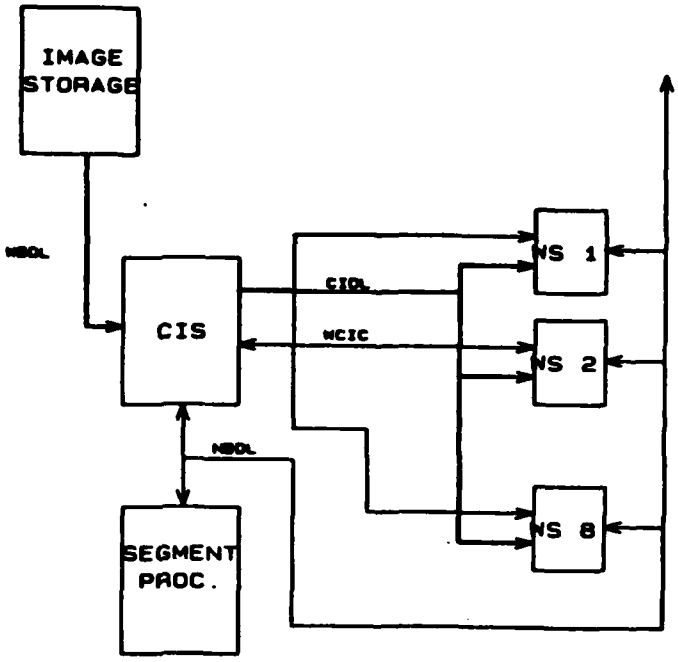
T0442(3)-002

REQUIREMENTS

- STORAGE
  - IMAGE DATA FOR EIGHT WORK STATIONS
  - 10.25 GIGA-BYTES
- DATA RATES
  - INPUT - 100 MBPS PEAK / 60 MBPS AVE.
  - OUTPUT - 2 SEC UPDATE RATE @ WORK STATION
- BER
  - $< 1 \times 10^{-10}$
- FAULT DETECTION
  - < 10 SECONDS

T0442(4)-003

CIS INTERFACES



INPUT / OUTPUT DATA RATES

● WIDEBAND DATA LINK (WBDL)	100 MBPS
● NARROWBAND DATA LINK (NBDL) (GFE)	156 KBPS
● CLUSTER IMAGE DATA LINK (CIDL)	360 MBPS
● WORK STATION TO CLUSTER IMAGE CONTROL (WCIC)	9600 BAUD

DATA INPUT / OUTPUT

- **INPUT**
  - IMAGE DATA
  - IMAGE TEST DATA
  - RECIRCULATED DATA
  - COMMAND AND CONTROL DATA
- **OUTPUT**
  - IMAGE DATA TO WORK STATION
  - IMAGE DATA TO WBDL (TBR)
  - COMMAND AND CONTROL DATA

T8502(4)-006

CIS CHARACTERISTICS

FORMATTED DATA CAPACITY	11,158 MBYTES
DATA TRANSFER RATE	
- PEAK	360 MBITS/SEC (10 MWPS)
- AVERAGE	273.7 MBITS/SEC (7.6 MWPS) (@ 36 BITS/WORD)
ACCESS TIME	
- AVERAGE	30 MILLISEC
- TRACK TO TRACK	5 MILLISEC
- MAX (EDGE TO EDGE)	52 MILLISEC
SIZE	ONE 5 FT. RACK
POWER	1.074 KWATTS
WEIGHT	< 500 LBS

T8502(4)-007

Contour Medical Systems
CEMAX 1000
Physicians Imaging Console

Contour Medical Systems
CEMAX-1000
Physicians Imaging Console
Specifications

Console
Intelligent, high resolution graphics system
—1024 × 1280
—19" color display
67 Megabyte digital cassette recorder
Digitizer tablet with cursor

Cabinet
9 Track magnetic tape drive
160 Megabyte Winchester disk drive
Modem

Host CPU
Multiple processor architecture
Dual 32-bit MC 68000's with cache memory

Image Display
256 shades of grey, color, or grey with color highlight
Formats
—Single
—Multiple up to 9 images
—Magnification

Image Processing
Universal image data input
Contrast Selection—window width, level, grey scale and color control
Window to window image copy, with or without magnification
Multiplanar Reformatting—sagittal, coronal, or oblique planes parallel to longitudinal axis
—Interactive selection of plane
—Maintenance of full CT number range
—Smoothing with interpolation
Variable axis selection

Data Base Management System
Automatic organization of patient directory from magnetic tape data
Creation of ProjectionView from magnetic tape data
Slice selection from ProjectionView or via menu
Temporary storage on Winchester disk
Permanent storage on digital tape
Powerful pre-editor

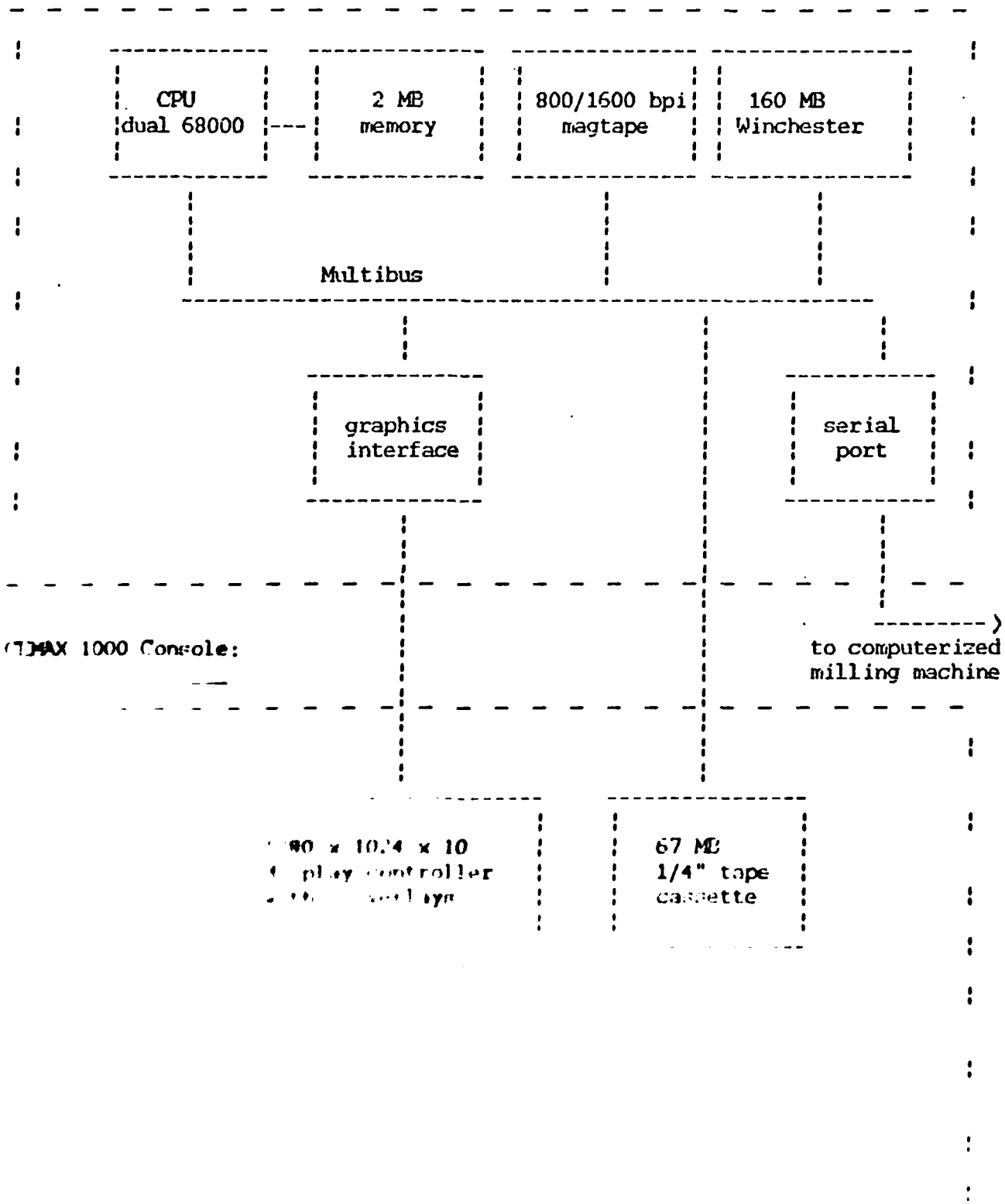
Image Analysis
Contouring
—Selection of tissue of interest by CT number
—Creation of volume of interest by specifying number of slices and region of interest within slice
—Automatic contouring of all tissue within volume and file storing on Winchester disk
—Option of smoothing data before contouring
—Editing and deletion of contours
Three dimensional views of contour
—Ring stack
—Shaded slices
—Shaded solid with DepthEncoding
—Shaded solid with Real Time Light Source
—Topographic
Volume measurement
—Volume calculation
—Deletion of contours

Future Options
Cephalometrics
—Two dimensional surgical simulation
—Soft tissue conforming
Networking compatibility
NMR image display and analysis
User definable protocols

Contour Medical Systems
1931-A Old Middlefield Road
Mountain View, California 94043
Telephone: (415) 969-2983

CEMAX - 1000 BLOCK DIAGRAM : HARDWARE

CEMAX-1000 Cabinet:



CEMAX-1000: SOFTWARE

Systems Software:

- Subset of UNIX Systems III, with real-time extensions.
- Drivers and libraries for high-resolution color display, digitizing tablet, cassette tape drive and numerically controlled milling machine.
- Intertask communication for windowed handling of tablet interrupt.
- Menu system generation and job control.
- Keyboardless user-interface.
- Local archiving to cassette tape.

Clinical Applications Software:

- Data transfer capability from archival magtape of numerous scanner manufacturers.
- Management of scan data and derived data on disk.
- Display, enhancement and measurement of slice data.
- Correction for movement and other artifact.
- Multiplanar reformation of slice data.
- Extraction of tissue surfaces by interpolation and density thresholding.
- Multiple types of display of three dimensional data including range-encoding, illumination models and transparency:
- Linear and volume measurements.
- Comparison of patient measurement with database of measurements (facial measurements).
- Editing tools to separate two three-dimensional objects, to modify them or to create new objects (e.g. implants).
- Generation of optimal tool paths for computer-controlled milling of molds or models.

ANALYSIS AND PROPOSAL

Parvati Dev

August 17, 1985

The following analysis and proposal is based on the Innsbruck proposal, Salazar's objectives document and group discussions.

1. There is an early need for a pilot machine on which brain mapping data can be processed. Use of this machine will initiate the necessary steps to standardize data representation, especially coordinate systems and nomenclature.

2. The pilot system must be "open," that is programmable. As one or all the laboratories, algorithms will be developed and tested for data processing feature extraction, display, etc. This development will guide the ongoing development of the Trixar modules.

3. The pilot system that is made available should make use of available experience. Don Woodward, Dean Hillman and Harvey Kerten has all developed data acquisition systems and display systems appropriate to brain mapping. Contour Medical Systems has developed an one vendor display system for most commercial CT and some MR scanners. It is currently implementing software to warp and register studies based on landmarks or surfaces. Fred Bookstein has considerable experience in deformations that relate sets of points. Toga has developed displays relating multiple studies. The pilot system should contain an integration of these capabilities.

4. The pilot system should have the best that is available in the [illegible] area. This [illegible] and the Science [illegible]. The [illegible] that would be readily [illegible] available [illegible] [illegible] [illegible].

5. [illegible] the [illegible] on the pilot [illegible]

the Trixar modules. A representation based on the concept of relational data bases and semantic sets is recommended. The entities in the database of the pilot machine are voxels or contours with the relation being the object or surface to which these entities belong. Note that the "object" in the pilot machine may be a neuron, a brain nucleus or a vessel. The resolution of levels of representation is left for later.

6. Figure 1 shows the necessary capabilities of a pilot machine. A modular, parallel development effort is proposed. It is assumed that off-the-shelf hardware will be used and that the necessary driver software for all peripherals will be available. Note that it is very important that each of these capabilities be extensible at each lab.

a. Specification of data representation including data types and coordinate systems. (E,F,G) Implementation of data base manager including a library of routines to read in, write from and query the data base.

b. Specification of data acquisition protocols. Implementation of modules for each type of data to be acquired. Each module can be developed independently as long as the output is in a standard voxel format with a standard coordinate system. (A,E,G)

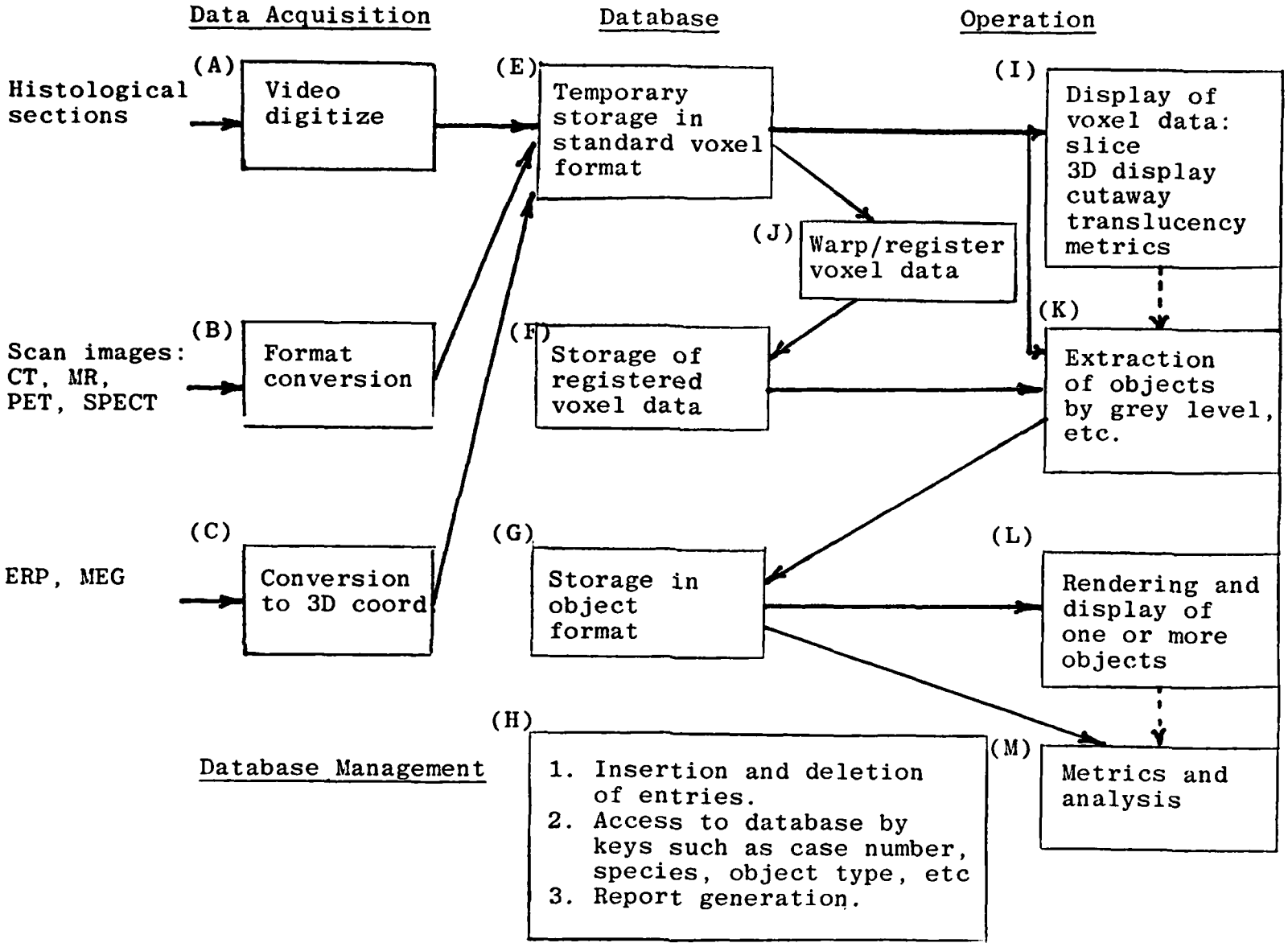
c. Specification of voxel display requirements. Implementation of a library of display routines and the ability to extend this library. (A)

d. Specification of real [illegible] and the necessary [illegible] interfaces. Implementation of the [illegible] vector library [illegible] and interfaces. (A)

e. Specification of some object extraction routines and of the means to add new algorithms, both automatic and interactive. Implementation of above. (H)

f. Specification of object display and data analysis requirements. This is what the user sees. Therefore, it will guide specification of a future brain mapping machine and hence must be very easily extensible. (L,H)

Figure 1


Data Acquisition
Database
Operation
Histological sections
(A)
Video digitize
(E)
Temporary storage in standard voxel format
(I)
Display of voxel data: slice 3D display cutaway translucency metrics
(J)
Warp/register voxel data
(K)
Scan images: CT, MR, PET, SPECT
(B)
Format conversion
(F)
Storage of registered voxel data
Extraction of objects by grey level, etc.
ERP, MEG
(C)
Conversion to 3D coord
(G)
Storage in object format
(L)
Rendering and display of one or more objects
(H)
Database Management
1. Insertion and deletion of entries.
2. Access to database by keys such as case number, species, object type, etc
3. Report generation.
(M)
Metrics and analysis

Interactive Solids Processing for Medical Analysis and Planning

Dr. Donald J. Meagher
Phoenix Data Systems, Inc.
80 Wolf Road
Albany, New York 12205

ABSTRACT

Valuable 3-D information concerning the internal condition of medical patients can be routinely acquired from Computed Tomography (CT) and Nuclear Magnetic Resonance (NMR) scanners. These scanners are limited to the display of individual 2-D slices, however, greatly reducing the usefulness of the data. The ability to effectively utilize such 3-D information is made possible by a new technique called "solids processing." Arbitrary 3-D solids are interactively manipulated, analyzed, and displayed for use in medical and other applications. A specialized hardware system based on a fundamentally new methodology was developed to eliminate problems in representation and performance. The applications of solids processing to diagnosis and treatment planning in craniofacial surgery is presented.

Introduction

The introduction of Computed Tomography (CT) has revolutionized several areas of medicine over the last decade. Detailed 3-D information on internal anatomical structures can now be routinely acquired without surgery. Similar improvements in other areas can be expected from the enhanced soft-tissue discrimination now becoming available with Nuclear Magnetic Resonance (NMR) scanners. NMR provides the additional benefit that the patient is not exposed to ionizing radiation.

Although a vast amount of 3-D information can now be easily collected, the ability to take full advantage of these data has not been similarly revolutionized. Radiologists and surgeons are typically presented with a series of 2-D slices on a CRT or in photographs.

A few computer programs have been developed to display 3-D medical objects derived from scanner information [1,2]. As shown in Figure 1, multiple 2-D slices are assembled into a 3-D array of volume elements, or "voxels." They are then viewed selectively according to density (CT number) or other characteristics. These programs have demonstrated the potential benefit that could be realized if medical objects derived from scanner data could be efficiently and effectively manipulated, analyzed, and displayed.

Such programs typically begin by generating a binary object (or a few objects) through density thresholding, manual region selection (on the 2-D slices) or both. One of three display methods is then employed.

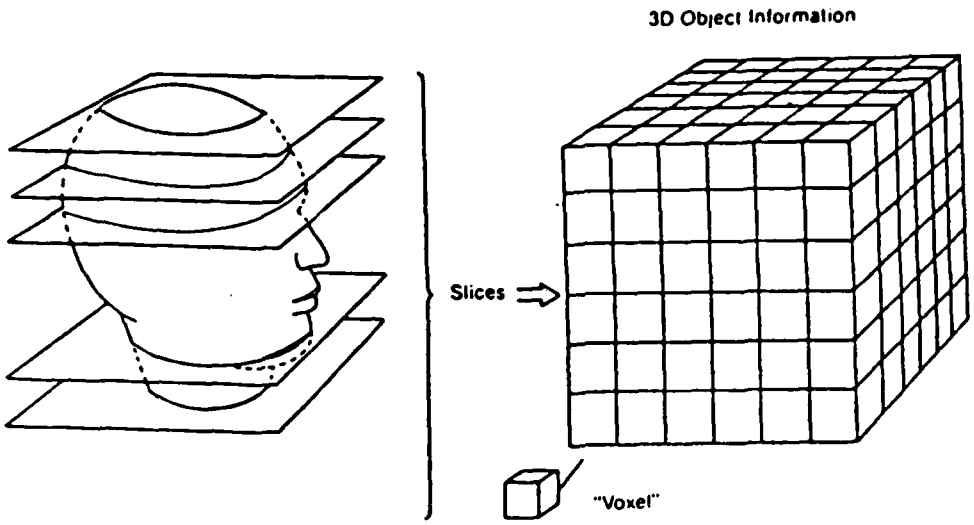


Figure 1 - Slices from a Medical Scanner are Assembled into a 3-D Array of "Voxels"

For the fastest response to user requests, the outer edges of the object are extracted from each slice and presented as a stack of contour lines on a high-performance 3-D line-drawing display system. Such displays are often used as part of large Computer-Aided Design (CAD) systems in the aircraft and automotive industries. Images are generated almost instantaneously in response to user requests. The main disadvantages of this method are the line display format (rather than surfaces of solid objects) and the lack of hidden-line removal. The images can become hopelessly confused as the number of contours increases.

The second display method extracts the surface faces of the individual voxels that make up the binary object, slice by slice. An image is then generated for a viewer in the plane of the slices, each slice corresponding to a horizontal band on a raster display. Viewer movement is restricted to rotation about the vertical axis of the object. The gray scale value given to each face element is typically based upon depth from display plane ("depth cueing") or some measure of tissue thickness previously extracted from the object. The presentation is somewhat more realistic than with the first method but usually requires a few tens of seconds to several minutes per image (the original surface extraction may take much longer than this).

The third display method extracts surface faces but then allows for the random movement of the viewer. More sophisticated surface shading techniques are often used. This provides more realistic looking objects but each image typically requires tens of minutes or longer to generate.

Scanner information could, of course, be used for much more than display or the simple extraction of measurements. In theory, just about any action or procedure that could be performed (or imagined) on the physical structures could be simulated in a computerized system. It should be possible, for example, to simulate a lengthy surgical procedure and view the expected results well before any surgery is actually performed. This opens up the possibility of evaluating various custom treatment strategies and plans for individual patients.

The main barrier prohibiting this type of use has been the excessive processing times required to manipulate and display such solid objects. It has simply not been possible to construct a system that could exhibit interactive performance. Interactivity, as used here, means a system response time of less than, say, a second or two for most user requests.

Why is interactivity necessary before such applications become viable? The answer lies in the nature of medical and surgical procedures: they are iterative. The results observed in previous steps are used to take the next step. During a surgical procedure, for example, an internal organ is examined closely before the type, location, and size of an incision is decided. In any computerized system, the results of past system requests would be used as a guide to determine the next request. If the response time is long, the user can lose his train of thought and even forget what he requested, perhaps rendering the system useless in many cases. On the other hand, the faster the response, the more iterations that can be performed. The user thus has a greater opportunity to examine and analyze the information and to investigate and evaluate more alternatives.

Solids Processing

There are three major reasons for the difficulty in developing interactive systems for processing solid objects derived from medical scanners. First, such systems often exhibit worse than linear growth in computations with object complexity because of the data structures and algorithms used. Second, large volumes of information are involved. Just accessing the objects in memory or from disk files can cause a considerable delay. Third, processing is slow because general-purpose computers rather than special-purpose processors are used. The first problem is, by far, the most serious, but all three must be solved to create a truly interactive system.

The system described below is the result of over six years of research and development devoted specifically to the solution of these problems [3]. The system represents a new technology for handling solids. New object representation methods, data structures, and algorithms were developed to solve the growth problems, and specialized machine architectures and hardware processors were designed to process the large volumes of information needed quickly.

The new technology is fully developed and commercially available. It is opening a new field called "solids processing" in medicine and in other areas where the generation, analysis, manipulation, and display of arbitrary 3-D solid objects is needed on an interactive basis.

Craniofacial Surgery

The first medical application of the new technology is in craniofacial surgery [4,5]. New diagnostic and surgical planning techniques are being developed in a joint effort of Phoenix Data Systems and Dr. David Hemmy, Department of Neurosurgery, Medical College of Wisconsin, Milwaukee.

Craniofacial surgery is performed to correct the sometimes grotesque deformities of the face and skull caused by accident, congenital abnormality, or genetic defect. This often involves major restructuring of the bones of the face and cranium, as well as modifications to soft tissue, such as relocating the eyes. Intricate and delicate procedures must be performed. Frequently,

bone grafts are used to fill gaps, modify surface contours, or fabricate missing sections. The bone is typically obtained from the surrounding area, or removed from the skull, ribs, or pelvic area. Parts of bone or even whole structures are detached and relocated. They are typically held in place with thin metal wires threaded through holes drilled by the surgeon or with metal brackets and small screws. Often, such procedures cannot be wholly decided on or planned prior to the operation, but must be customized during actual surgery, as information is gained. It is not unusual during an operation to discover that more bone is needed than was anticipated.

Because of the radical nature of the procedures, it is often necessary to violate traditional surgical principles. A certain amount of risk is involved. The relocation of an eye could result in blindness due to traction on the optic nerve. Movement of the brain could result in brain damage. The relocation of nerves could cause loss of hearing, speech, or feeling. Structural changes could result in air flow and breathing problems. In addition, relocated bones must be stablized in order to reestablish blood flow, without which they would disintegrate.

The need for precise structural information and detailed planning is clear. A multidisciplinary team approach has been found necessary. Neurosurgeons, plastic surgeons, anesthesiologists, and others (often including social workers and psychologists) investigate the deformity in detail, plan for safety, and document the treatment for later evaluation.

Most information is currently obtained from cephalometric radiographs. Unfortunately, in most situations it is of little use for analysis and planning. Years of relevant surgical experience are required to recognize the subtleties of the deformations. Usually sketches are made on full-size photographs and radiographs in an attempt to predict the changes needed and the results. Estimation of the size and shape of the parts involved, the amount of additional material needed and the best source for the material, the resulting soft tissue displacements, and the potential dangers are based upon very little information. New procedures are practiced on cadavers. It is believed that the analysis and planning process and the treatment results can be greatly improved if 3-D information is available in an effective and useful format before surgery. This includes visualization, measurement (distances, volumes, etc.) and the simulation of the procedures.

The Insight Medical Analysis and Planning System

The new technology for solids processing has been embodied in a specialized hardware processor called the Solids Engine™. A packaged system for stand-alone use in medical and other areas is shown in Figure 2. It is marketed under the name Insight™.

The block diagram of a typical configuration is presented in Figure 3. A central computer is used for communication, control, processor initialization, and application programs.

Figure 2 - The Insight™ Medical System

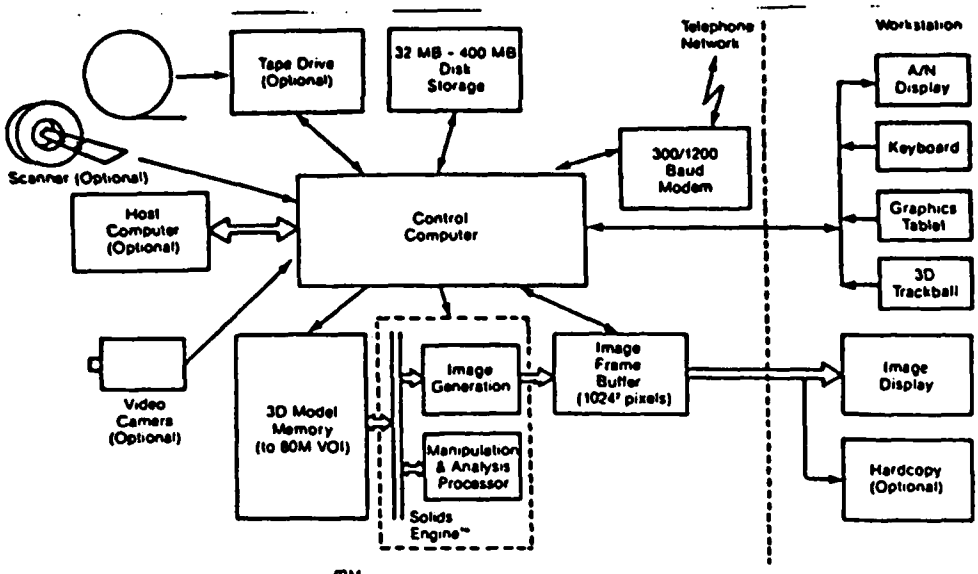


Figure 3 - The Insight™ System

Scanner information can be acquired directly from the scanner, via magnetic tape or through a host computer (RS-232, Ethernet, or direct connection). In some situations, information can be entered through a video camera (from existing hard copy, for example). Up to 400 MB of disk storage is available. A 300/1200 baud modem is used for external communications, messages, and remote diagnostics. The 3-D Model Memory is a high-speed memory that holds 3-D models. It is loaded through the computer and accessed by the Solids Engine . The Solids Engine is modular and can be divided into an image generation section and a section for manipulation and analysis.

The image generation is initialized by the computer in response to user requests. It then generates an image of the modeled object or objects and stores it in the Image Frame Buffer for viewing by the user. The Manipulation and Analysis Processor can be used to perform various operations on modeled objects.

The workstation can be located remote from the main processor cabinet. An alphanumeric display and keyboard are used for communicating with the internal computer or a host machine. Two interactive devices are provided, a graphics tablet and a 3-D trackball. Images are displayed on a raster CRT.

Object Acquisition

The progression from raw voxel to 3-D model is shown in Figure 4. Up to three density values per pixel can be acquired from the scanner (NMR information can be multidimensional). For one or two values per pixel, up to 16 bits each can be processed. For three values, up to 10 bits each can be used. A prefilter is used to eliminate unneeded voxels based on density. They are then compressed as images and stored on the disk.

The next step is to select the "voxels of interest" (VOI's) from the raw voxels. This is to eliminate extraneous tissues and structures that could obscure needed parts or cause distractions, and to reduce the total number of voxels that need to be stored in the 3-D Model Memory later. This is performed either automatically or interactively by a specialized processor, the VOI Selection Processor.

First the images are uncompressed and stored in the Image Frame Buffer memory. Several functions can then be randomly invoked by the user or under program control. One or more additional images can be interpolated between existing images. Various image processing operations can be performed such as filtering, density windowing, and, for region selection, edge detection.

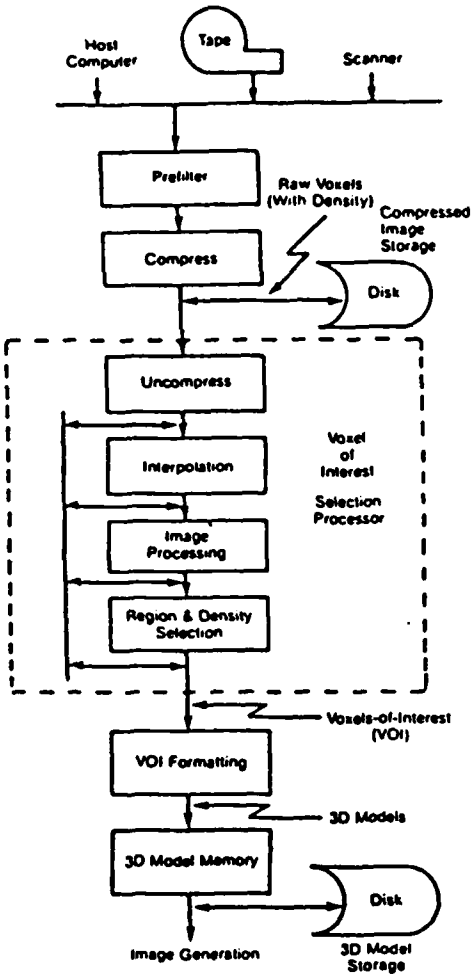


Figure 4 - Voxel Processing

Region and density selection are then performed. In some cases, such as selection of bone, simple high-pass density filtering can be performed. Other cases can be more difficult. Specific regions of the images can be selected in numerous ways. A rectangular box can be automatically or semi-automatically placed around the region of interest either on one slice for all images or for each image individually. The user can alternatively enclose a desired region by outlining it using the graphics tablet or by "painting" it using a variable size "paintbrush" attached to the trackball. Again, this can be done once or for each image. Depending on the exact situation, edge detection and connectivity analysis can facilitate this process. Density selection can be pre-determined based on past experience or interactively selected. Various density transformations can be performed to compress or enhance the information. The VOI selection process is important as a quick and easy way to remove obviously extraneous voxels. Several additional tools are later available in 3-D to complete the process, including multiple cut planes, 3-D connectivity analysis, dynamic density thresholding, and the use of the "electronic scalpel."

The resulting VOI's, complete with density information with each voxel, are then formatted into a 3-D model or models by a specialized processor. They can be directly loaded into the 3-D Model Memory for immediate display, manipulation, and analysis or stored on the disk for later use.

A voxel reduction of 10-to-1 is typical when selecting VOI's. In its maximum memory configuration, the 3-D Model Memory can hold up to 80M voxels of interest (the total of all objects being processed at the same time). Typical Insight configurations contain enough memory to hold a maximum of 12M VOI's.

A number of VOI storage formats can be used, depending on the number of density bits needed with each voxel. They are 1, 8, 16, 32, or 64 bits per voxel. The memory bus of the 3-D Model Memory was designed to be sufficiently wide (over 100 bits) that even 64 bits per voxel does not cause a substantial delay as compared to the smaller formats.

It should be noted that no information has been lost from the original scanner information except for those voxels specifically rejected or any loss due to deliberate density transformations. Neither the processing nor the storage causes any other reductions in the accuracy of precision of either density values or voxel locations. Also, the voxels can be randomly located in a universe of up to 64K by 64K by 64K voxels. There are no restrictions on object shape. Objects can involve multiple structures and tissues, they can be concave or disjoint, they can have interior voids, etc.

The two most commonly used formats for medical applications are the 1 bit format (for maximum object complexity) and the 16 bit format (maximum of 8M VOI's). The 1 bit format is typically used to differentiate between two subsets of the VOI's. They can be considered to be two objects, perhaps based on density. They can be made invisible, given random colors, or made translucent independently.

In the 16 bit format, 16 bits are stored for each voxel with the most significant 12 bits (4096 densities) used to determine visibility and color (pseudocolor). The remaining 4 bits can be used for possible density transformations on the object or for visibility determination based on which of 16 predetermined sets the voxel belongs to.

The upper 12 bits of a 16 bit density voxel are used to access two tables, the Pseudocolor Table and the Visibility Table. The Pseudocolor Table contains 4096 entries of 24 bits each (8 red, 8 green, and 8 blue). Each density can thus be displayed with a randomly selected color from a pallet of over 16M colors. The visibility table contains 4096 entries of 1 bit. Each specifies whether voxels with that density are to be visible or invisible.

The Pseudocolor Table and Visibility Table are loaded from computer memory into the Image Generator for each image generation cycle. Pseudocolor and visibility can thus be changed dramatically.

Display

Because of the need for interactivity, the display capability of Insight is of paramount importance. Most operations are built around the display. The user is given the ability to monitor and interact with the system dynamically. The system is normally controlled through menu items selected from the imaging CRT using the graphics tablet. Use of the alphanumeric CRT and the keyboard is minimized during an analysis and planning session.

A high-resolution (1024 by 1024) display is used for maximum image fidelity. Also, multiple image frames are used in order to make object movement as smooth as possible.

The display facility allows one or more objects to be displayed on one or more screen windows. The user can move dynamically to any viewpoint in 3-D space by rotating a 3-D trackball or selecting predetermined locations. The trackball can also be used to control additional display parameters. The image can be translated on the display screen. The objects can be scaled,

either uniformly, or independently in the three axes of the object coordinate systems. The 3-D location of the center of rotation of the object can be interactively moved. In all cases, three buttons on the trackball unit can be activated to lock out modification of the parameters attached to any of the rotation angles.

Hidden surfaces are removed and gray scale shaded surfaces are presented in all interactive solid display modes. Two shading methods are used: "block shading," in which voxels are displayed as solid blocks; and "surface-normal shading," in which local surface normals are used. Anti-aliasing is performed automatically by the Image Generator.

Up to three sets of two parallel cut planes (a total of 6 cut planes) can be manipulated. Each set can be randomly oriented in space and each plane can be translated independently. For each set of two, the user can elect to remove all material between the planes or, alternatively, all material outside the two planes. When multiple sets are in use, more sophisticated sectioning can be performed. Under user or program control, the system can independently display or inhibit display for up to 27 regions of space defined by the cut planes.

Depth cueing can be requested to make the surfaces of objects more dim as they become more distant from the observer. With Dynamic Density Thresholding (DDT), the user (or a program) can make specific densities or ranges of density visible or invisible (typically specified by CT number). When the pseudocolor feature is in operation, colors can be independently specified for each density. Also, objects or parts of objects can be made translucent so as to reveal interior details. This can be especially useful when comparing "before" and "after" objects

Analysis

The user can enter and [illegible] space. In addition, object surface points [illegible] and marked by rotating the object with the trackball [illegible] visible and then [illegible] means of point identification [illegible] attributes.

Once points in space [illegible] distances [illegible] operations are [illegible] fitted [illegible] entered [illegible]

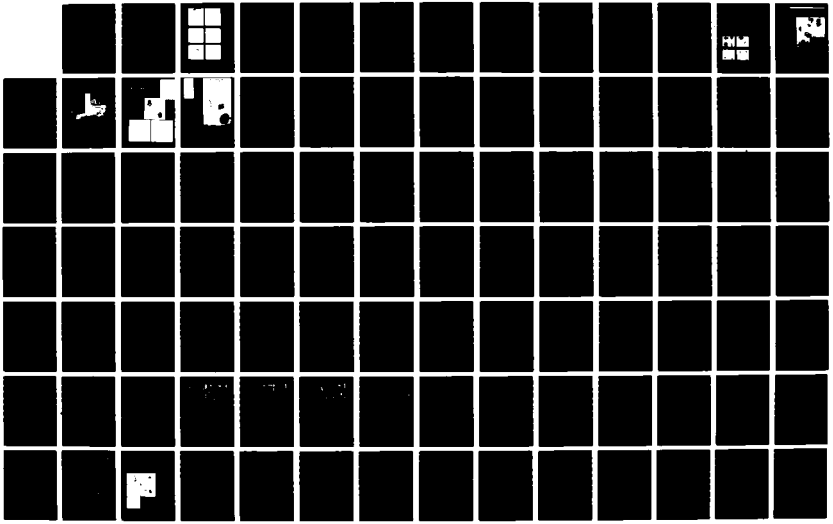
[illegible]

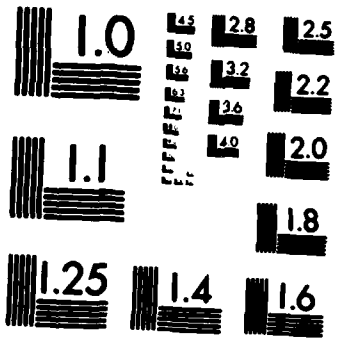
[illegible]

[illegible]

AD-A188 809 PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP HELD IN COLLEGE..(U) TEXAS A AND M UNIV COLLEGE STATION R B LIVINGSTON AUG 85 2/5

UNCLASSIFIED DAMD-17-85-G-5042 F/G 6/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

operations are used to translate, scale, rotate, skew, and reflect objects.

Two forms of connectivity analysis are possible. In the first, a surface "seed point" is selected as above. All voxels that are face connected to the seed point are then marked. The disjoint part can then be changed in color, made invisible (or everything else can be made invisible), removed, converted into a separate object, etc. In the second form of connectivity, all disjoint parts of an object are separated into unique objects. It is then possible to remove all parts below a specified volume automatically, retain only the largest or two largest parts, and so on. The number and size of interior voids can be determined and displayed.

A major feature of the Insight system is interactive interference detection. It can determine if an object has been moved into a location where it occupies space that is already occupied. It can, of course, prevent such situations by not allowing movement to be accepted if this would occur. This can be of great benefit when parts of bone, for example, are being relocated. The user is relieved of the difficult and often impossible task of "visual" interference detection.

The system has the ability to generate objects that are described mathematically. Various second and third order surface and solid objects can be produced. This allows the user or application programs to generate synthetic objects for auxiliary purposes. Small cylinders can be generated and subtracted from a piece of bone, for example, in order to simulate drill holes to be used for wiring.

A more advanced application is the "electronic scalpel," in which a "cut surface" can be interactively defined to separate an object into two.

The "time-varying" object capability can be used to switch between multiple objects with each image generated. This can be useful, for example, when several sets of CT scans of the heart are taken over a period of time, each with a different phase shift relative to the heartbeat. This results in a set of objects corresponding to the heart at a series of times during a pumping cycle. The operation of the heart can be displayed dynamically. All other functions such as density thresholding and cut planes operate as if the beating heart was a single object.

RESULTS

Insight units are in use at several major medical centers. Figure 5 shows a typical study of a slightly deformed skull of a child before corrective surgery was performed. It is based on 72 CT scans of 320 by 320 pixels each with a slice separation of 1.5 cm. No interpolated slices were used. The images are views from various locations in space. The orbits, lower jaw, and cranial cavity are clearly visible. Figure 5(b) was generated with a cut plane specified in order to expose the interior detail.

It should be noted that the object entering the mouth and extending down the throat is a tube inserted for medical purposes.

The images in Figure 5 were generated interactively with the viewpoint and cut plane locations randomly selected using the 3-D trackball. Such images are continuously generated. Most, including the ones shown, require less than one second between images.

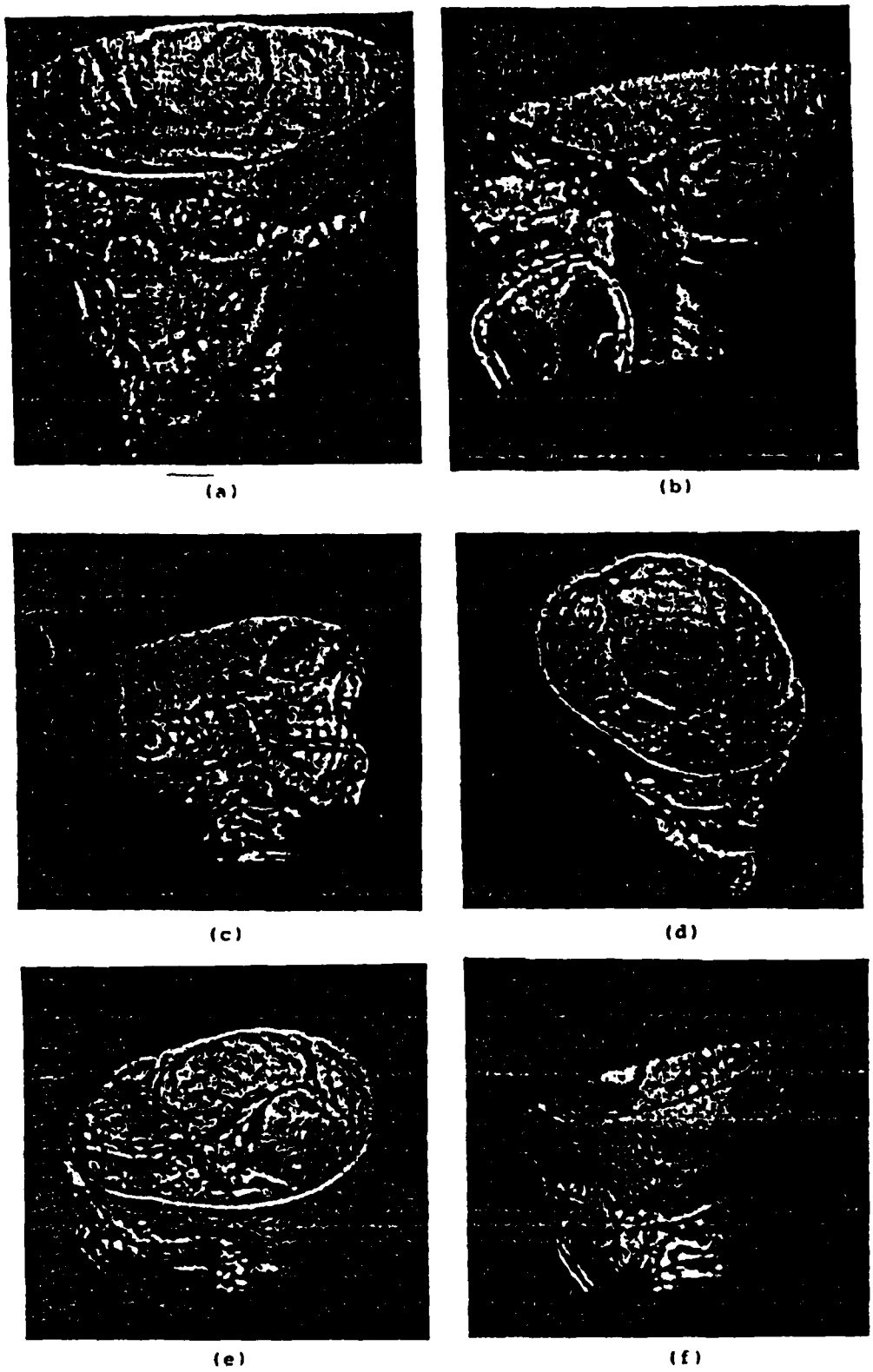


Figure 5 - Image of Human Skull

CONCLUSION

A fundamentally new solids processing capability has been developed. It has made dramatic new medical applications practical and is expected to lead to additional advances in the diagnosis and treatment of severe medical problems in the near future.

ACKNOWLEDGMENT

The author wishes to thank Dr. David C. Hemmy, Medical College of Wisconsin, for information used in the preparation of this paper and in Figure 5, and Ms. Brenda Ott for preparation of the manuscript.

REFERENCES

1. Artzy, E., Frieder, G., and Herman, G.T., "The Theory, Design, Implementation, and Evaluation of a Three-Dimensional Surface Detection Algorithm," Computer Graphics, Volume 14, Number 3, July, 1980.

2. Vannier, M.W., Marsh, J.L., and Warren, J.O., "Three Dimensional Computer Graphics for Craniofacial Surgical Planning and Evaluation," Computer Graphics, Volume 17, Number 3, July, 1983.

3. Meagher, D., "High Speed Display of 3-D Medical Images Using Octree Encoding," IPL-TR-021, Image Processing Laboratory, Rensselaer Polytechnic Institute, September, 1981.

4. Jackson, I.T., Munro, I.R., Salyer, K.E., and Whitaker, L.A., "Atlas of Craniomaxillofacial Surgery," The C.V. Mosby Company, 1982.

5. Hemmy, D.C., David, D.J., and Herman, G.T., "Three-Dimensional Reconstruction of Craniofacial Deformity Using Computed Tomography," Neurosurgery, Volume 13, Number 5, November, 1983.

A New Mathematics for Solids Processing

Octrees Make Widespread Applications a Reality

By Donald J. Meagher, Octree Machine Co. (formerly Phoenix Data Systems)

The two most serious problems in solids processing—the failure of processing to develop commensurately with object complexity and the limited domain of representable objects—have required the introduction of a new mathematics. The technique is based on eight-level hierarchical tree structures called "octrees," and it is the basis for a commercially available system that makes widespread application of solids processing possible.

Solids processing is the interactive manipulation, analysis, and display of computerized models of arbitrary solid objects by digital processors. Object models can be acquired from real-world objects or generated synthetically.

With a cost-effective solids processing system, manipulating and viewing solid models interactively becomes practical for many applications (e.g., medicine, CAD/CAM, AEC CAD, simulation, video games, molecular modeling, cinematography, publishing, and geophysical exploration) for the first time. The system allows practically any operation possible with physical objects to be performed with modeled objects. Anyone dealing with 3-D objects or information, or needing images of real or imagined objects or scenes, is a potential user.

Functionality

Experience has determined five solids processing functions to be useful in a large percentage of potential application areas:

☐ *Object Acquisition:* Models acquired from a variety of devices (e.g., CT and NMR scanners, electron microscopes, and laser scanners).

☐ *Object Generation:* Synthetic objects defined in a variety of input formats (e.g., polyhedra, second- and third-order surface and solid primitives), or the swept volume of any 2- or 3-D object.

☐ *Manipulation:* Geometric (e.g., translation, scaling, and rotation) and set (e.g., union, intersection, difference, and negation) operations.

☐ *Analysis:* Mass properties (e.g., surface area, volume, mass, center of gravity, and moment of inertia), static interference detection, dynamic interference (collision) detection, segmentation into disjoint parts, and finite-element mesh generation.

☐ *Display:* Display of any number of objects from any viewpoint, hidden surface removal, shaded surfaces, and cut planes.

All features should be unrestricted in use and interactive (i.e., with a typical response time of less than two seconds), even for complex objects and situations. The faster the response, the more iterations can be performed. Qualitative improvements will usually be the result if the designer has the opportunity to investigate and evaluate many alternatives. The user may also be able to produce more designs.

Object Representation Methods

The fundamental part of any system intended for solids processing is the internal object representation method employed. The two most popular methods have been boundary representation (B-Rep) and constructive solid geometry (CSG).

Reprinted from the October 1984 edition of Computer Graphics World

Binary Coordinate Axis

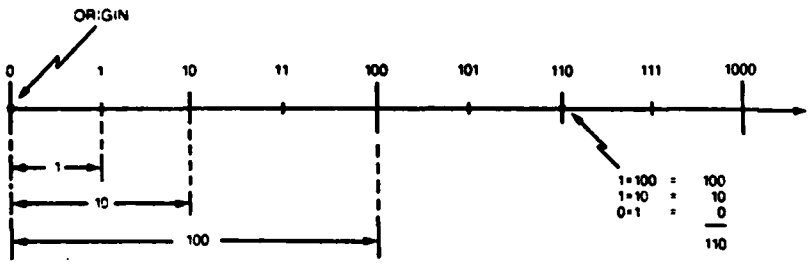


Figure 1a: Integer

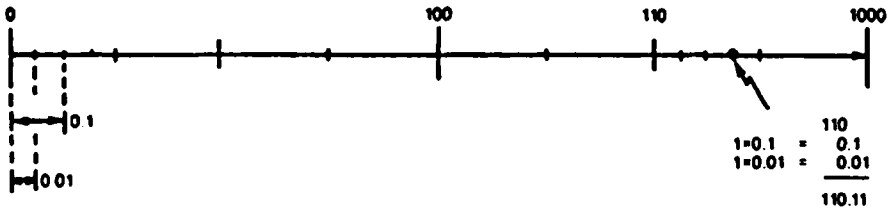


Figure 1b: Real

With B-Rep. solid objects are represented as the space enclosed by a collection of surface primitives. The surface elements may be planar or curved (usually limited to second order). A graph structure is typically used to maintain the object. Nodes corresponding to surfaces are connected to others representing the boundary edges which are, in turn, connected to nodes representing vertex points. In the CSG approach, solid rather than surface primitives are combined, using set operations to form objects. Tree structures are typically used, with leaf nodes corresponding to primitives and branch nodes representing set operations.

Two majors flaws in these methods have limited their use. First, the number of representable objects is very small. Modeled objects are typically formed from a library of not more than a few hundred simple "primitives," making the B-Rep method useless for many applications.

The second major flaw is performance. Because of the computational complexity of the data structures and algorithms employed, long processing times on large, expensive computers are usually required. Either long delays in system response must be tolerated (usually resulting in an unrealized productivity improvement), or a massive quantity of expensive, specialized hardware (e.g., flight simulators) must be employed to achieve fast response.

The performance problems are still not generally understood or appreciated. Common misconceptions hold that such problems are inherent to solid modeling and cannot be solved, or that performance would be greatly improved if the programs were rewritten to be more efficient or transferred to faster machines (or array processors).

The root problem is the often quadratic explosion of low-level elemental comparisons (e.g., spatial, sweep, and visual interferences) accompanying increases in object complexity. For example, a 100-fold increase in complexity can result in a 10,000-fold increase in processing time, depending upon the operation being performed. Such problems exist for most important solid modeling operations (e.g., set operations, mass property analysis, collision detection, and hidden surface removal) and can only be understood in light of a geometrical complexity analysis (a relatively new discipline within computer science) of the data structures and algorithms used in solid modeling.

Object Representation

Geometry and mathematics were revolutionized by the three-dimensional Cartesian coordinate system—the use of three coordinates (x,y,z) as reference points—which had practical applications in everything from road building to celestial mechanics. Algebra and geometry were united in analytic geometry, making new forms of analysis possible; science and engineering became vastly more productive.

Computers have obviously been very successful in dealing with the algebraic and Cartesian structures of traditional geometry and engineering in areas such as automated drafting. However, their very

foundations have the built-in limitations that three-dimensional symbols and their manipulations have to be translated, in almost all cases, to one-dimensional numerals and arithmetic operations. Even in advanced solid modeling systems, solid objects are not modeled by symbols representing solid elements, but by numbers and algebraic formulas in Cartesian space representing edges, vertices, enclosing surfaces, and the like. None of these elements are solid (enclosed space).

1-D Binary Interval Axis

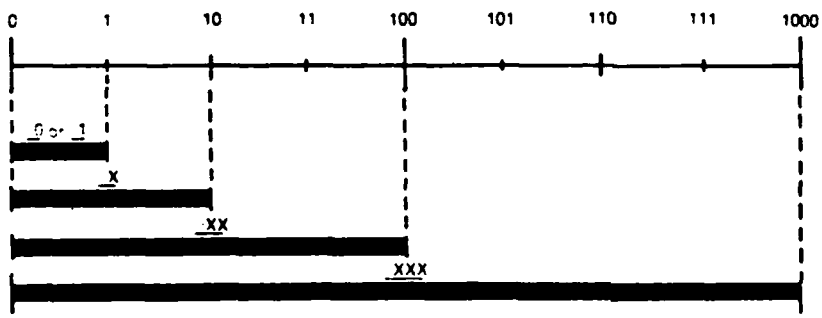


Figure 2a: Intervals

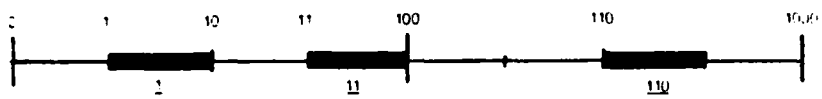


Figure 2b: Width 1

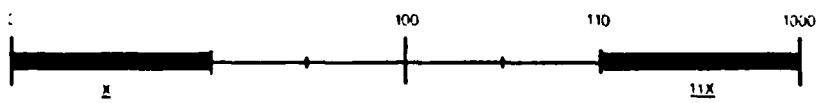


Figure 2c: Width 2

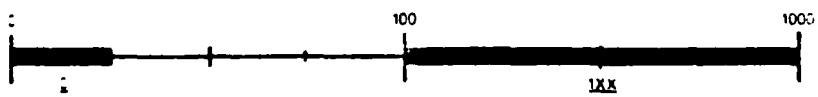


Figure 2d: Additional example

Figure 2e: Fractional interval

Geometric entities in Euclidean three-space are usually represented by points in a Cartesian coordinate system consisting of three orthogonal axes. They can be explicitly enumerated single points, or formulas, rules, or conventions defining sets of points (usually infinite sets). Thus, a line can be defined by an equation, and points for which the equation is true will be on the line. A line segment can be defined by two endpoints. Planes are similarly defined by equations, and polygons (regions of planes) by the boundary edges. This method of representation is extended in several ways for curved surfaces, and solids are defined by enclosing surfaces or formulas defining sets of points.

In the CSG approach, primitive solids are used. They are ultimately defined as the intersection of half-spaces (all the space on one side of a planar or curved surface that separates the space into two regions). The primitives themselves are then added, subtracted, and intersected to define the desired shape. The B-Rep method is somewhat similar, but segments of surfaces (not half-spaces) are "sewn" together (via common edges and vertices) to form a composite surface that encloses the space (i.e., separates interior from exterior volumes).

Object validity is a major problem with B-Rep. If any needed surface segment is missing, the interior and exterior are not separated and no solid object is defined. Later operations on such an "object" may

Bitree Representation of Segments

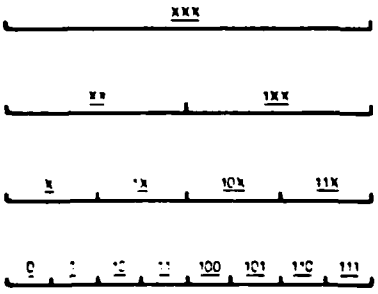


Figure 3a: Segments

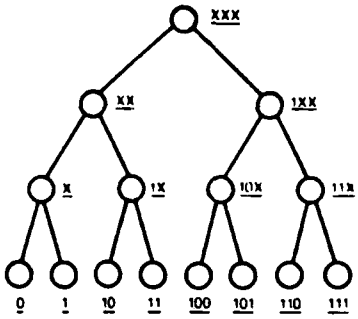


Figure 3b: Bitree

be meaningless. CSG does not have this problem.

Clearly, a B-Rep can be generated from a CSG object by locating all the exterior edges. This is, in fact, often done because operations such as display generation have been found to run faster in the B-Rep domain than in CSG. This is the "boundary evaluation" phase of many CSG-based systems. The evaluation itself takes a considerable amount of computation. It has a quadratic growth component because of the pairing of CSG primitives to determine if they intersect (and may, therefore, define an edge in the B-Rep).

Systems that handle solid objects using the above methods manipulate real and integral numbers corresponding to mathematical equations and data structures that define sets of points representing the solid objects. The new approach allows both the computer symbols and the corresponding internal electrical signals to represent three-dimensional space directly in a unified, single-dimensional coordinate system.

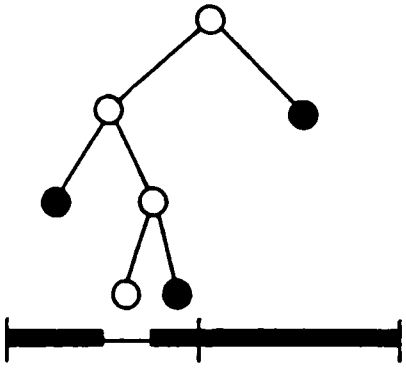


Figure 4: Sample bitree

The distinction between objects and signals would, perhaps, be academic were it not for what some consider to be a "fatal flaw" in the conventional approach: the computer's inability to sort solid objects and parts of solid objects spatially. Being essentially unrestricted in three dimensions, computers cannot rank objects in a one-dimensional space. They cannot even be ranked according to a one-dimensional projection because they usually overlap randomly and cannot, therefore, be represented by a single coordinate on a number line.

This inability to maintain objects in a pre-sorted form leads to sorting, searching, and comparison operations to fulfill individual user requests during system operation. In a display operation, for example, a different visual-priority sorting of polygons may be needed for each new viewpoint to remove hidden surfaces. If individual elements are not somehow sorted, pairing may be necessary even for comparisons, the most fundamental operations performed in such situations. The result of the inherently quadratic relationship between object complexity and required computations is a system in which transformations get much slower as object complexity increases. One can just imagine the state computerization would be in if sorting were not possible.

A New Approach to Object Representation

In most computer systems, points in 3-D space are represented by a set of three coordinates. For the sake of simplicity, one coordinate axis is considered in *Figure 1a*. The number line is labeled with binary numbers which are multiples of an arbitrary unit distance. Points are identified by the sum of the individual binary digits associated with each integer. The point at 110, for example, is displaced by 110 (1 × 100 + 1 × 10 + 0 × 1) binary units from the origin.

This is extended in *Figure 1b* to include not only unlimited length but also unlimited precision. Real numbers are defined by introducing a binary point. Digits to the right of the binary point are, of course, weighted by fractions of the

Quadtree

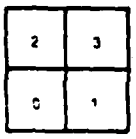


Figure 5a

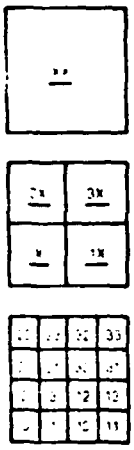
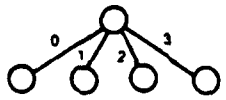


Figure 5b



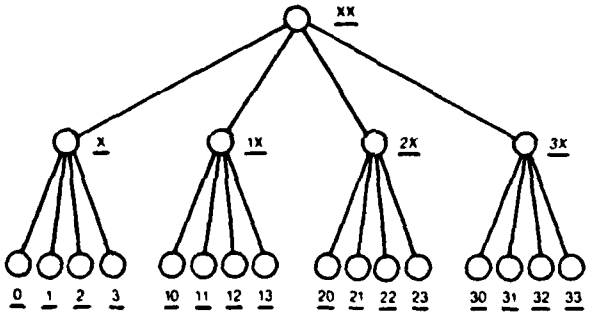


Figure 5c

unit distance. The point 0.11 (1 × .1 + 1 × .01) beyond location 110 is noted.

The concept of a point on a line is extended to represent "segments" (sections of lines enclosing distance). *Figure 2* presents a "binary segment" axis. Segments are represented by single numbers, underscored to distinguish them from points. The widths of individual segments are integral multiples of the smallest resolvable unit of distance. The value of the segment number is the location of the minimum point on the segment (closest to the origin).

In addition to the binary symbols 0 and 1, a third symbol "X" is used to define zero for determining the width of a segment. An X indicates that the segment extends across the boundary corresponding to that digit location. Thus, in *Figure 2a*, if the rightmost digit is a 0 or 1, the segment is one unit distance in width. If the rightmost digit is X (and no others), the segment is two units; if the two rightmost, 4 units; and so on. Any X's are located as a block in the least significant digits, and can be thought of as "place holders."

Figure 2b is an example. The segment *1* begins at location 1 and is of unit width. Segments *11* and *110* are also shown. In *Figure 2c*, two double-width segments are shown (the rightmost digit is X). In *Figure 2d*, the segment *0* is of unit width and is located at the origin. Segment *1XX* is four units in width and located at 100. The definition is easily extended to real-valued segments. Fractional segments are shown in *Figure 2e*. The number of digits to the right of the binary point indicates the width of the segment. The X symbol is not needed.

The above representation method is very limited in that it can represent only a single segment, the location is restricted, and the width can only be the minimum width to the second power. It is extended to represent more complex 1-D "objects" by combining multiple intervals in a specialized data structure called a "bitree." Nodes in a hierarchical tree represent the segments. For example, the eight unit-wide segments in *Figure 3a* are represented by the eight lowest level leaf nodes in *3b*. The segments of width two are represented by the parent nodes at the next highest level in the tree. Parent generation continues until the largest segment (the "universe") is represented by the root node of the bitree. Fractional segments could, of course, be represented by additional levels below the level of the "segment point." Clearly, all possible segments are represented.

Nodes are marked black only if they are part of an object. White nodes represent segments that either partly intersect the object or are entirely disjoint from the object. Of particular significance is the fact that all nodes below a black node or below a white node representing a disjoint segment are redundant and can be eliminated.

A sample bitree is presented in *Figure 4*. The object consists of two disjoint segments. The left part is two units in width; the right, five units. The root node segment partly intersects the object and is white. The right child node is entirely enclosed and is black. Although it has no children, the left node contains two. The left child contains the left

Sample Quadtree

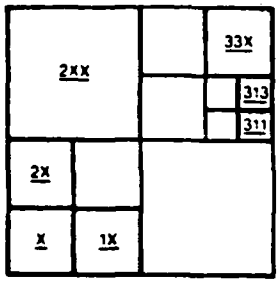


Figure 6a: 2-D object

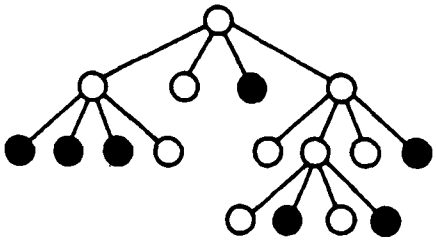


Figure 6b: Corresponding quadtree

Octree

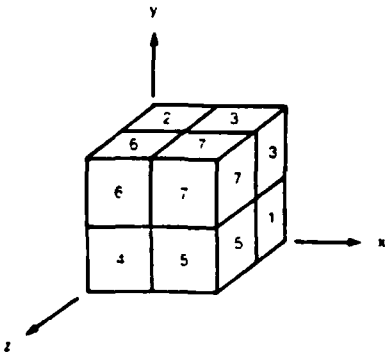
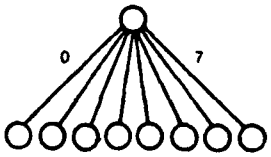


Figure 7a



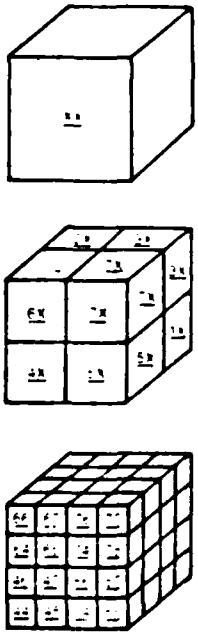


Figure 7b

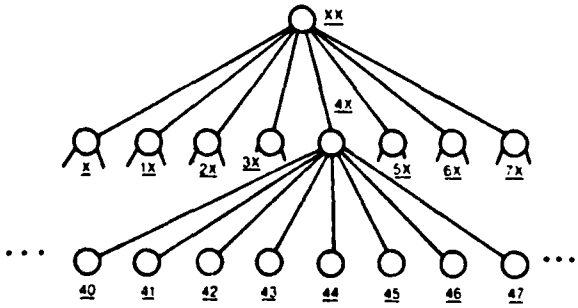


Figure 7c

part of the object and is black; the right one contains the two additional nodes to complete the object.

This method is extended into 2-D with the "quadtree." As shown in *Figure 5a*, each branch node contains four children rather than two. They represent the four quadrants of a parent square, and their numbers are shown in *Figure 5b*. The corresponding quadtree is in *5c*. A sample 2-D object and its quadtree is shown in *Figures 6a* and *b*.

For 3-D solid objects, the encoding is extended into one more dimension. An "octree" is defined in which each branch node contains eight children. Octant numbering is presented in *Figure 7a*. The cubical regions of space represented by nodes will be called "octants." Fractional octants exist below the level of the "octant point."

In *Figure 7b*, octants in a 4 by 4 by 4 universe are shown with octant numbering. The corresponding octree is shown in *7c*. An object consisting of three bottom-level cubes is shown in *Figure 8*. The universe is represented by the root node. Children 4 and 5 contain the object and are continued at the next level. The others are white leaf nodes. At the next level, three black nodes form the object.

Octree Encoding

The octree data structure allows arbitrary objects to be encoded to a variable precision, which can vary even within a single object, depending on the lowest tree level used for different parts of the object, or within local regions of space. The location of octants, relative to the origin of the universe, is exact. Any inaccuracy that exists because of the approximate nature of the en-

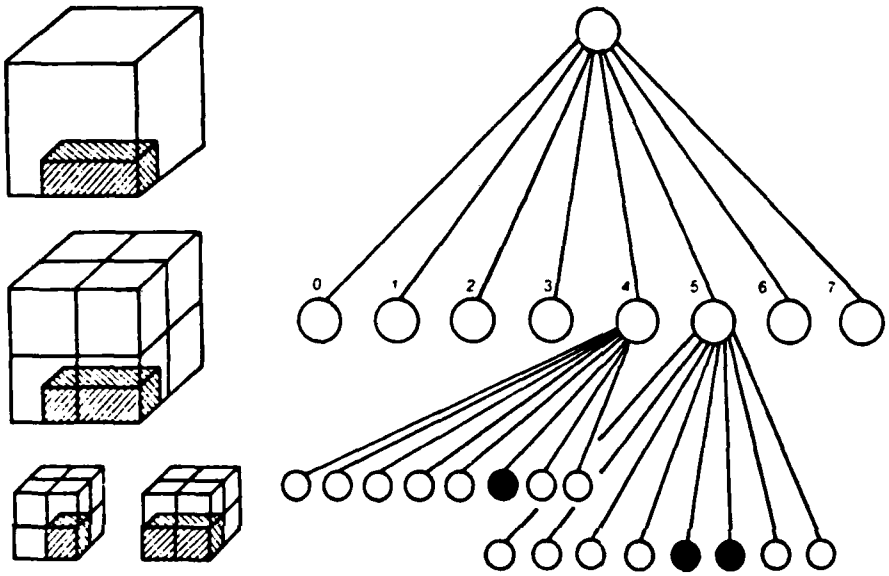


Figure 8: Sample quadtree

coding results from determining surface points within lowest level octants, not between octants.

It can be shown that the number of nodes needed to represent an object is on the order of the surface area of the object divided by the square of the resolution (resolution being the length of an edge of an octant at the lowest level used). The number of nodes needed for representation can be used as a measure of object complexity.

In most cases, octrees tend to be wide rather than deep (the number of levels is usually limited). For an octree with a maximum of 32 levels, for example, a resolution of 0.001 inch results in a universe enclosing 311,482.8 cubic miles.

The octant number is the concatenation of the individual child numbers that must be selected to traverse from the root to a particular node (padded with X's, if not at the unit octant level). Thus, the location of any octant is determined by a single number within which all dimensions are interleaved. The number can be used to sort octants spatially. A sorted list of octant numbers (one for each solid octant) is known as a "linear octree." When the octree itself is used, multiple sortings are maintained, each corresponding to different recursive traversal sequences.

Just as with any useful symbol system, operations on the symbol structures must be defined and algorithms developed. Operations on octrees and the associated algorithms have been extensively developed in recent years.

Point inclusion (determining if a specified point is interior or exterior to an object) is the most fundamental operation needed when dealing with solids. With octrees, this simply involves traversing down the tree, beginning at the root, always selecting the child octant containing the point. When a leaf node is reached, its type (black or white) specifies the status of the point. Points on common edges or faces are handled by defining them as a single octant or by testing multiple octants.

Set operations are performed by simultaneously traversing two or more octrees while generating an additional one. Nodes examined represent the same regions of space in all trees. Nodes in the output tree are generated based on the node types found in the input trees. Because the trees are examined in a sorted order, no quadratic growth pairings are needed.

Geometric operations are performed by traversing the original octree to determine the state of all octants that spatially intersect nodes being generated for the transformed octree. Some restricted operations, such as 90- and 180-degree rotations, can be performed by simply reordering nodes.

The calculation of mass property values is straightforward for octree objects. Algorithms have been developed for converting random sets and arrays of voxels (volume elements) into octrees and for efficiently generating octrees from CSG and B-Rep objects.

The octree structure greatly facilitates the removal of hidden surfaces in display generation. Depending on the location of the viewer, there exists an easily determined traversal sequence that, if applied recursively, will visit solid octants in such a way that no octant later in the sequence can obscure an octant visited earlier. If these nodes are projected onto a display screen with later nodes fill-

ing previously undefined pixels, an image with hidden surfaces removed is produced. In advanced versions of the algorithm, computations are related to the visual complexity of the image generated rather than the complexity of the objects viewed.

Additional algorithms have been developed for swept volume generation, static interference detection, dynamic interference detection (collision avoidance), finite-element modeling, and cross-correlation for machine vision. ■

Octree Implementation

Octrees and octree-like methods have recently been embodied in solids processing products for stand-alone application and are now being used at select industrial and medical sites. They have already demonstrated levels of performance several orders of magnitude greater than previously possible, with images being generated at approximately one million nodes per second.

Figure 9a shows four views of a human skull generated from 72 CT scans of 320 × 320 pixels each. The octree representing the object contains approximately 600,000 nodes (object extending down throat is a plastic tube). *Figure 9b* illustrates the use of a cutplane to reveal interior detail.

In the union of a cuboid and a discus-shaped object presented in *Figure 10*, cutplanes once again remove part of the object. Interior nodes have been rendered invisible, leaving only the surface.

The use of a "surface decal" is demonstrated in *Figure 11*. Properties are attached to individual nodes. In this case, the 24-bit color values from the ubiquitous Mandrill picture are mapped onto a turbine blade.

These images were generated interactively with the viewpoint and cutplane locations randomly selected using a 3-D trackball. Such images are continuously generated and usually require less than one second between images. ■

Courtesy of Dr. David C. Hemmy, Medical College of Wisconsin

Figure 9a: Four views of skull generated from 72 CT scans.

Courtesy of Dr. David C. Hemmy, Medical College of Wisconsin

Figure 9b: Cutplane reveals interior detail of skull.

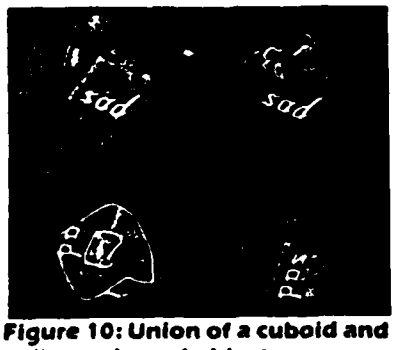


Figure 10: Union of a cuboid and a discus-shaped object.

Courtesy of Dr. Michael Potmesil, Bell Labs.

Figure 11: 24-bit color values from Mandrill image mapped onto turbine blade demonstrates "surface decal."

References

Hunter, G. M. "Efficient Computation and Data Structures for Graphics." Ph.D. dissertation. Electrical Engineering and Computer Science Department, Princeton University, June 1978.

Hunter, G. M., and K. Steiglitz. "Operations on Images Using Quad Trees." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, April 1979.

Jackins, C. L., and S. L. Tanimoto. "Oct-Trees and Their Use in Representing Three-Dimensional Objects." *Computer Graphics and Image Processing*, December 1980.

Meagher, D. "Octree Encoding: A New Technique for the Representation, Manipulation and Display of Arbitrary 3-D Objects by Computer." *Technical Report IPL-TR-80-111*, Image Processing Laboratory, Rensselaer Polytechnic Institute, October 1980.

Meagher, D. "Octree Generation, Analysis and Manipulation." *Technical Report IPL-TR-027*, Image Processing Laboratory, Rensselaer Polytechnic Institute, April 1982.

Meagher, D. "Efficient Synthetic Image Generation of Arbitrary 3-D Objects." *Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing*, June 1982.

Meagher, D. "Geometric Modeling Using Octree Encoding." *Computer Graphics and Image Processing*, 19 (June 1982).

Meagher, D. "Computer Software for Robotic Vision." Society of Photo-Optical Instrumentation Engineers' 26th Technical Symposium, August 1982.

Meagher, D. "The Octree Encoding Method for Efficient Solid Modeling." Ph.D. dissertation. Electrical, Computer and Systems Engineering Department, Rensselaer Polytechnic Institute, August 1982.

The author wishes to thank Professor Herbert Freeman, Image Processing Laboratory, Rensselaer Polytechnic Institute, for guidance in the octree development effort.

This article is an adaptation of a paper presented at Computer Graphics '84 and is published here with the kind permission of the National Computer Graphics Association.

Donald Meagher is the director of computer graphics development at Phoenix Data Systems (Albany, NY), where he has been active in the development of new methods and processors for handling solids.

Dr. Meagher also has worked for RPI, Argo Systems, Calspan, and IBM. He holds a B.S., M.Eng. and Ph.D. in Electrical Engineering and an M.Eng. in Computer and Systems Engineering from RPI.

Technology

SPEEDING UP THE REVOLUTION IN 3-D COMPUTER-AIDED DESIGN

NEW SYSTEMS CAN CREATE CHANGEABLE, LIFELIKE IMAGES IN A FRACTION OF A SECOND

Office workers grow extremely impatient when a computer takes longer than a second or two to display an answer to a problem. But their irritation would really soar if they had to wait for a computer-aided design system to generate three-dimensional images of solid objects and display them on a video screen. Not too long ago, this task could take hours to do, even with a large computer. "After you made a change [in the design] and wanted to see a display, you could go have lunch," recalls Herbert B. Voelcker, professor of electrical engineering at the University of Rochester and director of its Production Automation Project.

The time that it took computers to perform such jobs almost scuttled the early market for computer-aided design systems that do 3-D modeling. Now, thanks to new and faster ways to "paint" video pictures, computerized 3-D modeling has a good chance to move back on the fast-growth track the industry expected when the technology first appeared seven years ago. A handful of young companies recently has developed innovative "engines"—special-purpose computers—that can display these video pictures in a matter of seconds.

BREAKTHROUGH. For two of the new systems—from Phoenix Data Systems Inc. in Albany, N.Y., and Silicon Graphics Inc. in Mountain View, Calif.—painting the screen takes little more than an eyeblink. "The speed is just awesome—a breakthrough," says Thomas M. Rafferty, director of CAD marketing and development for McDonnell Douglas Automation Co. (McAuto), which is evaluating a Phoenix Data prototype. "It's going to become the expected thing for the whole CAD/CAM market."

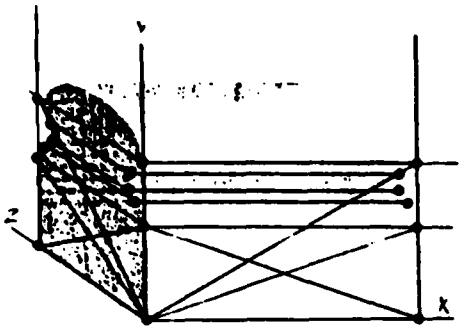
The ability to create 3-D models of solid objects, not just the line drawings or surface representations to which CAD systems were formerly limited, is the key to automated manufacturing. "This is of immense significance to the whole issue of America's industrial productivity," says Carl Machover, a CAD consultant and technical director of the Special Interest Group on Computer Graphics, a unit of the Association for Computing Machinery. If these systems provide an effective solution to the CAD imaging problem, he says, they will go a long way toward linking CAD and computer-aided manufacturing (CAM), so that a model developed on a CAD system can be used to program numerically controlled machining centers and factory robots.

NEUROSURGEON DAVID HEMMY (RIGHT) USES A COMPUTER SYSTEM DEVELOPED BY ALVIN RING'S PHOENIX DATA TO STUDY A 3-D MODEL OF A SPINAL SECTION

The superfast systems also have promise in applications other than designing products on a computer screen. One early possibility is "artificial vision systems for robots," says Rafferty of McAuto. "We're working closely with our robotics people on that." The enormous number-crunching capabilities of the new image processors also promise to push 3-D modeling beyond manufacturing into such large-scale applications as air-traffic control, weather prediction, seismic prospecting for oil, and molecular and genetic engineering.

Medicine may be one of the most exciting applications. One of Phoenix Data's prototype systems is being used by Dr. David C. Hemmy, associate professor of neurosurgery at the Medical College of Wisconsin, to perform realistic practice operations on the head and spinal cord before a real patient goes under the knife. Until now, performing a simulated operation on a 3-D display would have been almost unthinkable—for the same reason that solids modeling has been a disappointment to the CAD business. After any kind of change is made on the screen, even just rotating the image slightly, a system takes 1 to 5 minutes to update the display.

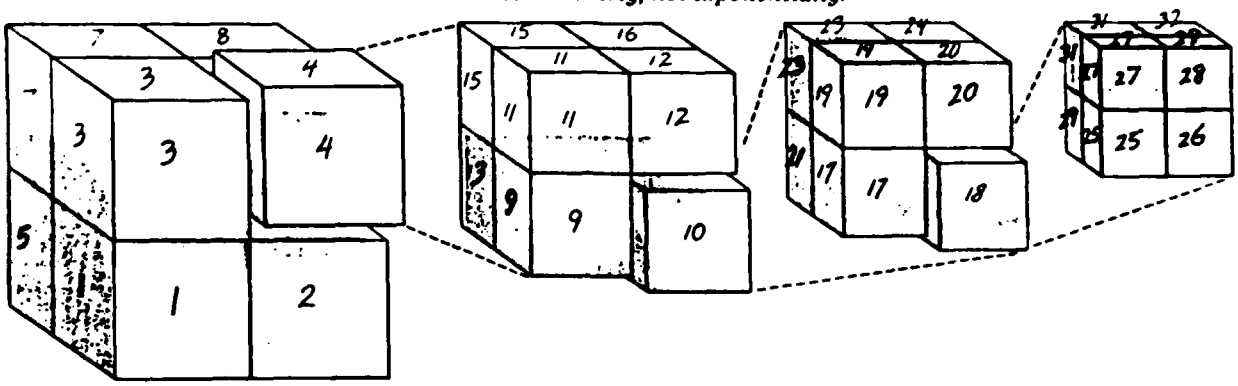
LIFE-SAVING. Dr. Hemmy begins by generating a 3-D model of the area to be operated on from computer-aided tomography (CAT scanners). Data from the two-dimensional CAT scans, taken every 5 mm of thickness, are fed into an image generator that stacks the slices to form a solid model. Then, seated at the screen, Dr. Hemmy uses a "trackball" controller to simulate, say, the reconstruction of a crushed jaw. He can cut through the simulated tissue and remove bone fragments, insert bone grafts and stainless-steel support pins, and then ask the computer to check to make sure that the additions do not in-



A FASTER WAY TO DISPLAY SOLID SHAPES

THE CURRENT APPROACH *(left)*
A problem with computer-aided design systems has been the time it takes to create video images of solid objects. Such three-dimensional images are built up from thousands of interconnected points and lines. To change or shift the display, this enormous mass of data is recalculated, which can take up to five minutes or longer because the number of calculations quadruples as the amount of data doubles.

HOW PHOENIX DATA DOES IT *(below)*
Phoenix Data's processor reverses this approach: It begins with the largest chunks of the image and adds a collection of progressively smaller cubes. First the image is enclosed within eight cubes. The cubes are then sorted: Those filled by the object are saved in memory, empty cubes are discarded, and partly filled cubes are again divided and sorted. This is repeated until only filled cubes remain. Each cube has its own location code, eliminating the need for x-y-z coordinates. This method can update the image in a fraction of a second, since the calculations required increase linearly, not exponentially.



terfere with the patient's circulation. Such practice operations, says Dr. Hemmy, will "help assure that the patient receives the right operation and cut down on the length of the surgery, significantly reducing the morbidity and even the mortality associated with a lot of surgical procedures."

Picker International would like to provide similar capabilities to hospitals that have bought its CAT scanners and is now evaluating Phoenix Data's equipment. And for nuclear magnetic resonance (NMR) machines, which make images similar to CAT scans but eliminate the danger of X-ray radiation, Picker may build in Phoenix's imaging system. For NMR, says Donald E. Plante, Picker's vice-president of engineering, the Phoenix technology "offers fantastic potential." General Electric Co.'s Medical Systems Business Div., which also markets body scanners, is working with a similar 3-D imaging engine from Weitek Corp. in Sunnyvale, Calif.

Solids modeling has been slow to catch on primarily because the CAD equipment makers failed to anticipate how much computation would be needed to provide a display that responds instantaneously, says Martin D. Schussel, director of mechanical software for Schlumberger Ltd.'s Applicon Inc. This happened, he says, because most models used for experiments [illegible] when you get up to a reasonable size model—something that somebody really wants to work with—the interactivity disappears very quickly."

The problem is that the mathematical descriptions in a model contain an enormous number of elements: points, lines, and arcs, plus equations that specify how these elements are interconnected in 3-D space. These have to be translated onto the video screen, a 2-D grid made up of about 1 million picture elements, or pixels. That task is relatively simple for a few elements, but as the number of variables doubles, the number of calculations needed to do the job quadruples. Even a supercomputer can take hours to make a change in a complex model and update the display.

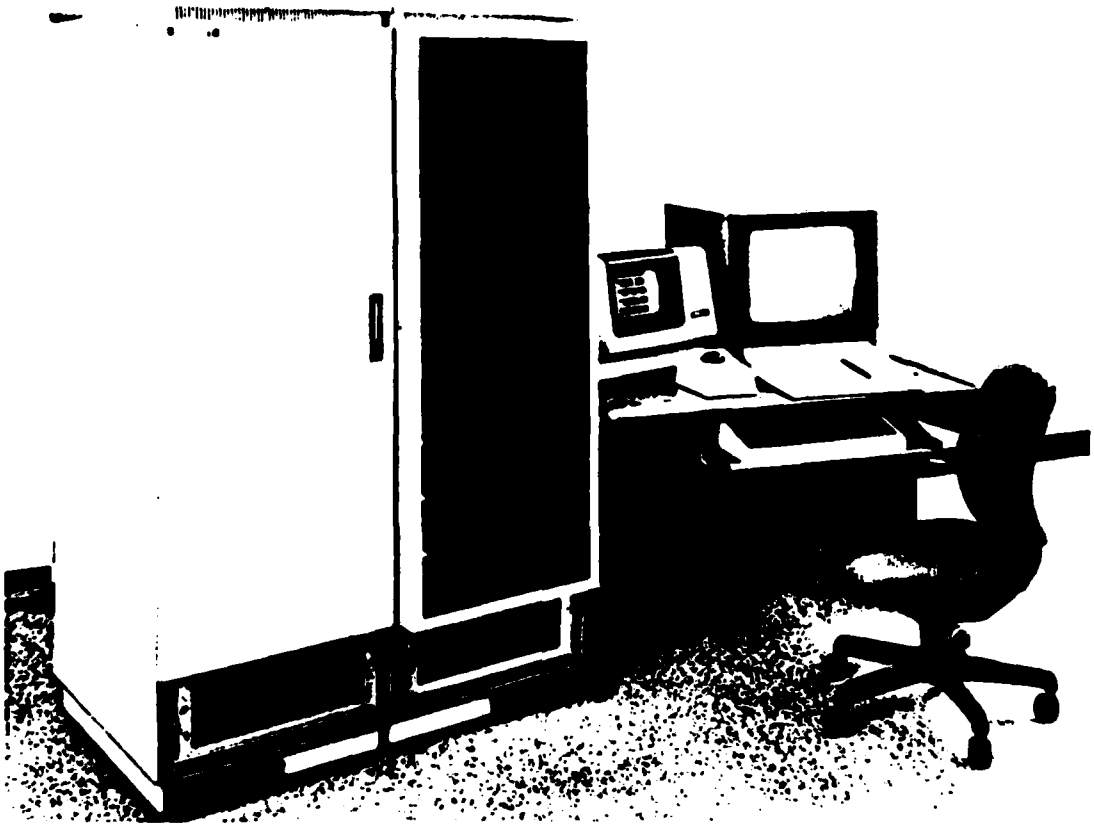
"It's like the old puzzle about how many grains of corn it takes to fill up a checkerboard if you double the number of grains on each successive square," notes Machover. "By the time you get to the 64th square, you need more grains than there are in the world."

DIVIDE AND SORT. To get around this problem, the new image processors rely on clever ways of sorting data and proprietary computer designs specifically tailored to handle 3-D data. Phoenix's image generator, for example, has eight computers that operate in parallel to divide a 3-D model into progressively smaller sets of eight cubes, sorting them into three categories: those filled by the model are saved in memory, the empty ones, which are discarded, and those that are partly filled. Any partly full cube is redivided and sorted until only full cubes remain (drawing).

The formula for the divide-and-sort operation was devised by Donald J. Meagher, computer graphics development director at Phoenix. "The algorithm is very unusual," notes Phoenix President Alvin J. Ring, "in that it procedes in a strictly linear manner"—that is, it never gets bogged down in an exponential-growth situation. As a result, screen changes occur in roughly one-tenth of a second.

Philip R. Kennicott, a researcher at General Electric Co.'s Automation & Control Laboratory, thinks Meagher's technique could be the breakthrough that CAD users have been waiting for. "It's a very powerful approach to solids modeling," he says. The image generator built around Meagher's algorithm, adds Rochester University's Voelcker, is "the fastest imaging system around."

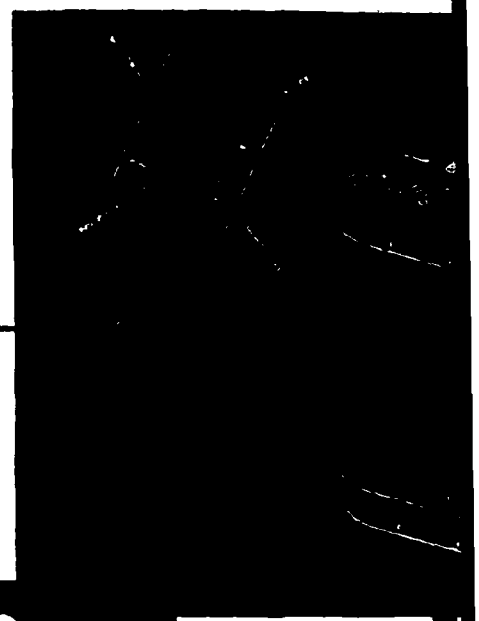
Most of the other high-speed imaging systems, he says, are variations of what the industry calls tiling engines. These contenders fabricate a display from "tiles" plastered across the surfaces of a model. The locations of the tiles are sorted in a way that if a background tile is obscured by a foreground tile, the system simply ignores it. Tiling engines are usually able to paint a new screen in less than 5 seconds, as a result. Says Voelcker: "I think we've finally entered the era where you can display things in fancy ways—and very quickly." ■



INSIGHT, the computer graphics system from Phoenix Data Systems of Albany, New York, generates true three-dimensional objects which can be displayed, manipulated and analyzed in real time. Other systems take hours to perform tasks that take INSIGHT only seconds.

INSIGHT provides full capability in the configuration shown above, which includes the Solids Engine™ (housed in the cabinet) and the interactive workstation tool. A low-cost Solids Engine™ option, suitable for CAD/CAM and other applications, is now available.

Reprinted from Electronic Imaging,
March, 1985



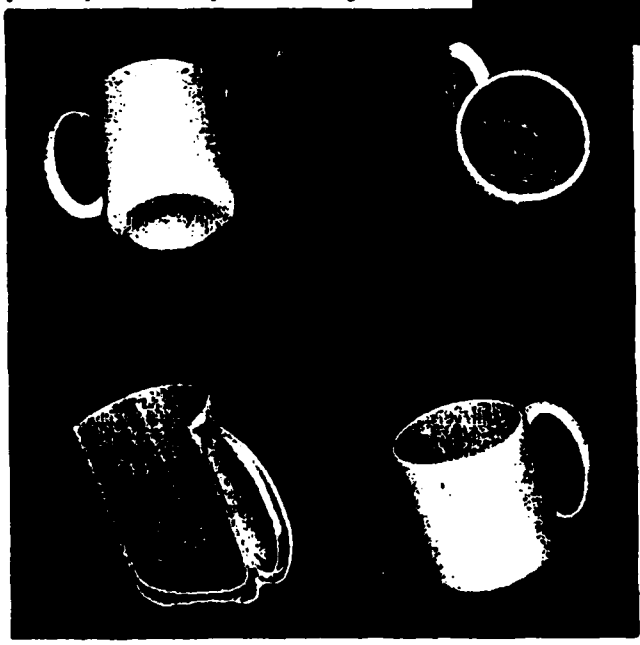
News Update

Solids engine enhanced for imaging

Processors to be implemented on the Solids Engine 3D graphics system from Phoenix Data Systems (Albany, NY) will expand its range of applications from mainly medical areas at present to lower cost uses in CAD/CAM and other markets. The system is currently configured with a display processor for image generation and a second processor for set operations and interference detection.

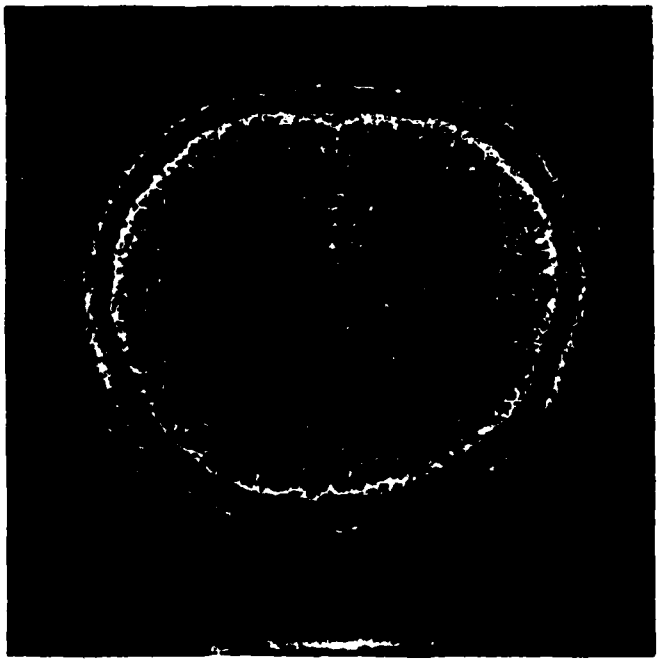
The company's Insight medical analysis and planning system incorporates this hardware on a custom modular bus structure called Object-bus, running at 40 million nodes per second. Three dimensional volume elements or "voxels" are created from 2D scanner images and then processed to form 3D graphic solid models.

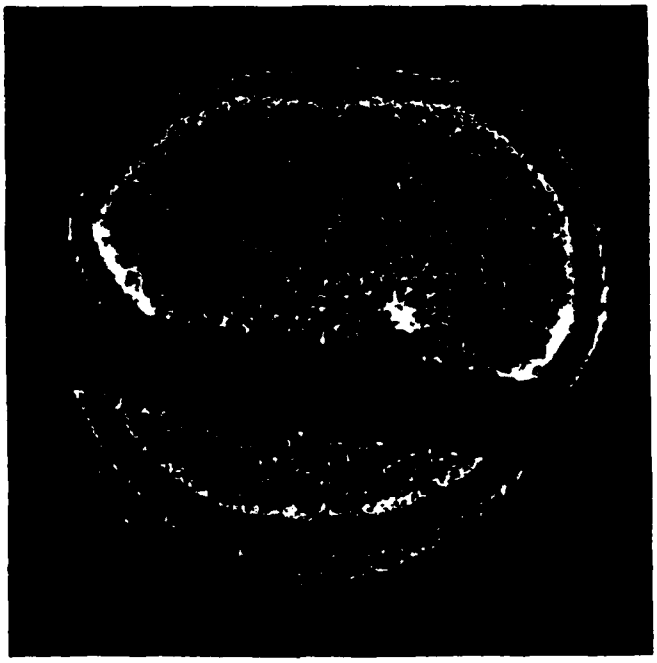
The Insight system can be used to compare 3D graphics of implants with CT scan images. To make a prosthetic replacement hip joint, for example, approximately 50 parameters must be known in order to specify the implant sufficiently accurately, says Donald Meagher of Phoenix. An average of 70 CT scans can be built up to form a 3D graphic model on the Insight system. This model is then compared with the 50 joint parameters using the interference detection capability which determines automatically whether the two match up adequately for the implant operation to proceed using that



Solids engine hardware additions allow more than just medical uses.

Brain tumor studies at the M.D. Anderson Tumor Clinic in Houston, TX, use 3D data from CT scans (below left).
Dynamic density thresholding software reveals the tumor (below).





Hardware implementation of processing functions will extend applications into CAD/CAM.

replacement joint shape.

Meagher is now working on new processors for the Object-bus, to be discussed at the NCGA conference next month. The aim is to incorporate software image processing features into hardware.

The first new processor performs geometric operations such as linear transformation, rotation and scaling, with collision detection between displayed objects. The existing graphics generation hardware can perform these changes of viewpoint but cannot detect collisions.

The second processor, called CSG-1, for continuous solid geometry, is used to request second-order solid primitives. By defining the primitives using a combination of second-order and planar half-spaces, says Meagher, it is possible to generate relatively complex shapes such as pipes rather than being limited to simple cylinders, spheres and ellipsoids.

Software is currently available on the Solids Engine for the computation of surface area, volume, mass, center of mass and moment of inertia of displayed solid models. The software is used in medical applications, for example to locate the center of mass of a tumor so that a radiation source may be inserted at the best point to destroy the tumor.

Hardware implementation of the mass properties processor will add new application areas, says Meagher. It will enable engineers to calculate the weights of parts in mechanical design or to know the moment of inertia of a satellite design, for example. Software currently under development at Phoenix is the forerunner of a CSG-2 hardware processor, and will be used to generate swept volumes in space for robotic applications.

Information can be obtained from Donald Meagher at (518) 159 / 202

2. Neuroanatomical Data Acquisition, Analysis and Display

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619 455-9100

EMMA Research System Developed by Scripps Clinic and Research Foundation 2.1.1

by Floyd Bloom

August 9, 1985

Dr. Robert Livingston
Prof of Neurosciences
Dept of Neurosciences, M-024
UCSD
La Jolla, CA 92093

Dear Bob:

This letter and accompanying materials will attempt to set down on paper the basic philosophy and current operational status of the research instruments we are building here for the scientific analysis of neuroanatomic structures. Warren Young and I will come to the conference to provide whatever personal input we can offer for the full working day of Tuesday Aug 13 and we appreciate your understanding of the pressures with respect to our upcoming major Site Visit.

In brief, the device we are assembling is primarily a research system in three parts, as illustrated within. The central element is "EMMA", our acronym for Electronic Morphomtery and Mapping Analysis, built around a top of the line Zeiss Optical Microscope with two non-standard components: 1) Highly accurate, passive, position sensors on the X, Y axes of the stage, and on the Z (focus) axis); 2) A high resolution video camera attached to an extra image viewing port to provide either a "raw" video image or a digitized video image. The other two legs of our three part device are the on-line Atlas of brain structures (one for rat, one for cynomolgus), and a "Data Base" of neuroanatomic structures (again, one for rat and one for monkey). The objectives are to provide the user with the ability to search microscopic specimens in planes approximating the standard atlas planes, identify the region, area, nucleus, layer or cell set that they have marked with their experimental or cytochemical procedures, and then link their observations to what is available within the literature on that site.

Some of the major features of our system are: 1) the overt intention to use high resolution video (for multiple observers) as an acceptable alternative to analysis (one person at a time) through the microscope's objectives; 2) user-friendly interfaces to reduce to a minimum the need for conventional computer keyboard entries; 3) the ability to use a single system for a variety

of mapping, and morphometric measurements in which the software is relatively affordable, and the programming is user modifiable; 4) the ability to change magnifications and quantify cell density and distribution across areas larger than a single microscopic field. Other advantageous features may be revealed by a bit more detailed description of the three elements, and the existing analytic subroutines already envisioned, and in many cases written, and in use.

The user of the system can "map" the locations of cells, neuropil, or any other marked structural elements onto an in-memory template composed of the three planes of any page from the digitized pertinent reference electronic brain stereotaxic atlas. The user can determine the location of the structural elements to be mapped by reference to any of 4 atlas pages and zoom in on a particular atlas field to match the magnification of the microscopic section being studied. The user can also set off landmarks for a particular section, mapping the elements in a 2-dimensional array, and store that set to "stack" onto a serial section containing referenceable superimposable landmarks. (We are specifically designing our 2-dimensional data acquisition and storage routines to be compatible with the 3-D analysis systems developed by our colleague Dr. Arthur Olsen of the Molecular Biology Department. An ultimate goal, will be to define the locations of marked places in a 3-dimensional matrix, such that the matrices can be compared quantitatively for labelled structures). The user can also trace individual cell structures, or accurately measure any individual neuronal details (dendritic branches, cell sizes, axon diameters, etc).

The Atlas system is an integral accompaniment of the microscope for use in the locations of labelled structures, while the user is on-line or off-line. The Atlas also furnishes blank templates onto which to record the information of new mapped systems, and allows for the comparison, qualitatively, of two different data sets. We intend to develop programs that will under user direction examine the features of a set of structural elements (axon width, varicosity width, intervaricosity distance, etc), and from the digitized densities of such structures, provide a quantitative symbolic graphic representation onto the atlas format that will described the density and distribution properties of a specific cell marker. We believe that such data are currently not employed routinely because they are so hard to acquire and compare, but that these values can be attained with a system like ours.

The third element of our system is a searchable data base of neuroanatomic structures defined in terms of specific locations, their known afferent and efferent systems, and whatever is known regarding transmitters and other systems markers, and cellular or behavioral function. This system is organized in terms of a hierarchical array of nested organized brain structures (see organization outline enclosed) such that one can search on large structures and gain information on all nested related structures, depending on the level to which

data are avilable or the users' questions are definable. For example, the system could describe all of the afferents to a given structure, to a given layer of that structure, or to a specific type of neuron within a specified layer of a specified structure, depending upon what is known and declarable.

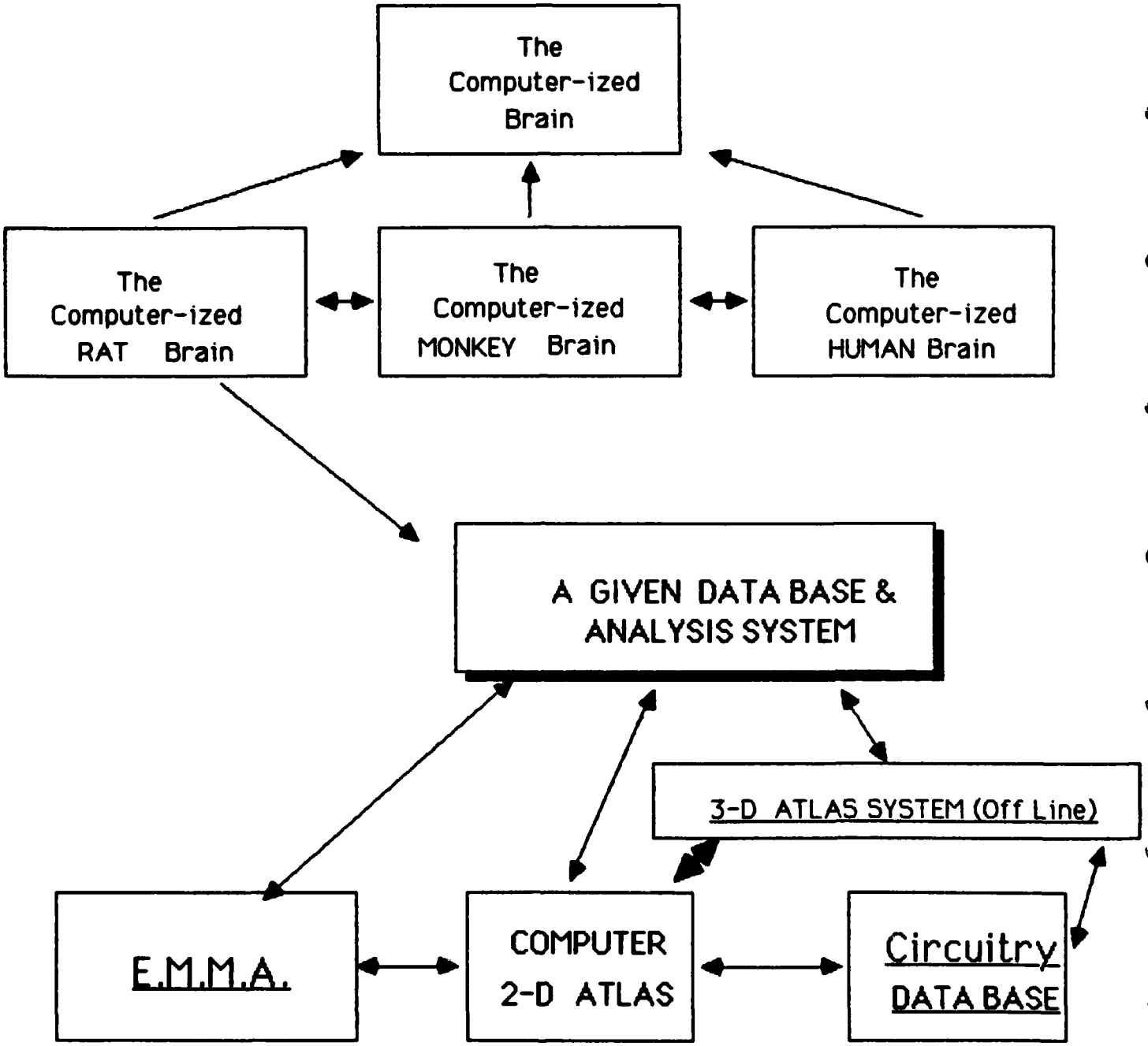
We see the system being implemented first for rat and cynomolgus monkey brains. Even with those beginning data bases and atlases, we recognize that we must also develop a translational system to compare data from one brain to another. Currently, we speculate that we may be able to do so by using the organized structural outlines of place names as a lexicon for the translations. We see a similar process as approaching the level of translatability that will allow the development Mega-System you are seeking to build. To approach the human brain in 3-D with all of the above intrinsic cellular, circuitry, and chemical data. We need to be able to incorporate meaningfully cellular and circuitry data from experimental animals. At least in my view, it is not likely ever to be possible that we will be able to gain for the human nervous system the types of details that experimental brains can furnish in terms connectivity, and good neurochemical information.

If the above was clear, then the enclosed detailed descriptions should be even more self-explanatory, and if not, that is what Warren and I can try to do when we arrive on Tuesday.

With best regards

Floyd

Floyd E. Bloom, M.D.
Director
Div of Preclinical Neuroscience
and Endocrinology



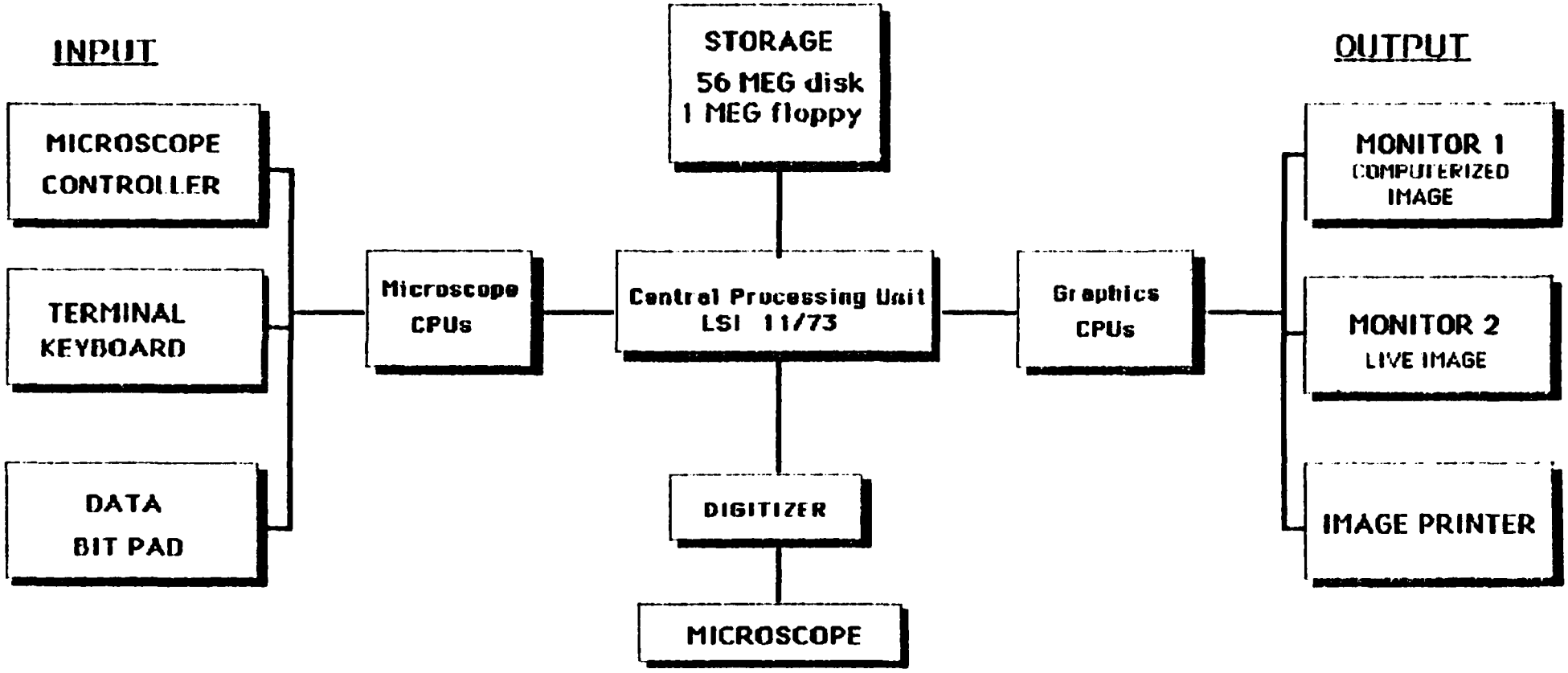
DATA SOURCE

Brain Sections | Atlases, Published Data | System Definitions

USES

Mapping | Structural Comparisons | Species Comparisons

Quantitative Morphometry



SCRIPPS EMMA

ELECTRONIC MORPHOMETRY and MAPPING ANALYSIS SYSTEM

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619455-9100

A description of EMMA and the ATLAS parts of this system follows:

Computer Hardware

The imaging system uses a Digital Equipment Corporation (DEC) LSI 11/73 central processing unit (CPU) on a DEC Q-bus. It has presently two and a quarter megabytes of main memory, with 58 megabytes of storage capacity. Several communication ports in the computer support multiple computer terminal setups as well as telephone communication. DEC RT-11 is the operating system, with TSX-Plus as the time share system. RT-11 is the fastest operating system available from DEC, designed specifically to support real time data acquisition. TSX-Plus supports a maximum of 31 concurrent users. The main LSI 11/73 CPU communicates with peripheral processors that make up the remainder of our imaging system: 1) the graphics display unit, 2) the microscope position sensors, 3) the electronic camera interfaces, and 4) the electronic bitpad.

Graphics Display Unit

The graphics display unit (GDU), Vectrix model VX384A, has a display resolution of 672 by 480 pixels. The image is 9 bitplanes deep, giving each pixel the capacity to display one of 512 separate colors or shades of gray. The GDU drives either an analog red, green, blue (RGB) monitor, a monochrome monitor, a color ink-jet printer, or a monochrome laser printer. Besides displaying the image, the GDU also contains the important program routines that create simple geometric shapes. These graphic primitives draw the lines, dots, and polygons in a two or three dimension space, and coordinate their spatial relationships during translations, rotations, and scalings. The importance of these GDU routines is to remove some of the computing time and burden from the CPU, greatly increasing the speed of the imaging system. All of the primitives are also implemented in more sophisticated routines (for example, display a picture file on the GDU) and are maintained in disk files called libraries. Since these routines are generic, program development time is significantly reduced.

Microscope Position Sensors

In many applications, the imaging system must keep track of the absolute positioning of the microscope stage in order to render positional coordinates of the neural structures. This positional information may track a cell fiber across large distances. It can also be used to relate different neural structures together during analysis. It must be capable of resolution at least on the micron level, with ranges on the order of centimeters. To achieve this goal, we have machined our Zeiss inverted microscope to accept Burleigh optical position sensors (Model CE-2000). The sensors operate on an optical fringe method, are compensated for thermal expansion, yield a real spatial precision of 1 micron, and can be interpolated to a tenth of a micron. Future expansion of

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619 455-9100

our system may include active, programmable position translators which piggy-back onto the sensors. These translators can then move the stage in either of the three axes through piezoelectric crystal movements. With the optical sensors, repeatability and accuracy of placement are in the micron region. All microscope stage devices are controlled by a second peripheral processor. This also serves the same important purpose of lowering the main CPU burden.

Electronic Camera and Image Digitizer Interface

All of the images from the Zeiss microscope are captured by a Cohu 5000 electronic camera, equipped with a Newvicon imaging tube. This Newvicon tube has a horizontal resolution of 900 lines and is capable of reading in low-light situations as during fluorescent studies. The image may be viewed directly on a high-resolution monochrome monitor or be stored in the computer for later recall and analysis.

If it is stored in the computer system, the image is first converted or digitized to a three dimensional matrix that is, in essence, a mathematical copy of the live image. A Colorado Video Digitizer (CVI 270A-1) does this work under control of the CPU. It can provide a resolution of up to 2048 horizontal units, keeping up with the camera at the very least and providing for future expansion to higher resolution cameras. The digitized image is 8 bitplanes deep, meaning that each point in the image can be from 1 to 256 levels of gray, or color. Once stored in a computer file, the image may be redisplayed at any time without any loss in quality (important for labile images such as fluorescence).

In many cases, the live image, displayed on the monitor, functions without digitization as a background image onto which computer drawn graphics are overlaid. This is useful during morphometric anaylsis when tracing structures.

Electronic Bitpad

The imaging programs get user feedback through two methods. One is the computer terminal keyboard. This is appropriate when a lot of characters have to be entered in response to program needs. However, it is fairly difficult, if not impossible to enter information that pertains to graphics and disrupts attention to anatomic details portrayed on the monitor. For example, the movement of a cursor on the GDU monitor to trace a fiber would be awkward from the keyboard. Hence, an electronic bitpad with an electronic mouse (Summagraphics model MM1201) serves as a major source of user input. Programs use the bitpad in different ways. One application is the display of menus and commands on the GDU monitor whereby selection of commands is initiated by pointing the cursor (by moving the mouse over the surface of the bitpad) at the selection and pressing a button on the mouse. The CPU can read these responses and take the correct

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619455-9100

action. Also, as the mouse movement can be very smooth in any direction, fibers can be traced easily with the bitpad. Most of our imaging programs that deal with data acquisition use the bitpad and mouse as the main user input device, freeing the user from the keyboard, and premitting more mobility during the work session.

Software

Most of the programs of our imaging system were written in Pascal, this particular variant conforming mostly to the protocol set by the International Standards Organization. Since Pascal is a high level, highly structured language, program development is relatively easy, even with the complex data manipulation normally found in imaging applications. Pascal is also fairly portable, whereby programs in our system can be carried and installed on other computer systems with only minor modifications. The Macro Assembler is used to create machine level code for time-critical routines that require very fast device responses. The machine code is still directly controlled through the smarter Pascal programs.

The programs in the imaging system follow a strict protocol. They are all user-friendly. Error and status messages are abundant if inappropriate responses are made. The programs always attempts to correct the errors made by the user. They also fully protect him against errors (accidentally deleting or overwriting existing files) or crashing the system. To reduce user error further all programs are controlled through a single computer monitor, the User Interface (UI). UI provides for the selection of programs through combinations of six different keys on the keyboard. The different programs are indexed as topics in system categories. Thus, there is a system for dealing with the microscope, one for the atlases, another for the digitizer. The user simply enters the system level using the keyboard arrow keys to point to the system. The actual program names need not be remembered by anyone. Help (brief text descriptions or answers to common problems) is always on available each topic and UI will display this help on the terminal screen.

Each program also is responsible for checking the validity of the data files that the user specifies. This is important because different system levels have different functions, and these functions or signatures are carried into the data files. In simplest terms, it is virtually impossible to analyze a data file that has incorrect data in it, even if the user makes an error and tries to analyze that incorrect file. The second purpose of this strict file protocol is more subtle. The internal organization of the data files follows a conventional structure used by many graphics systems. We adhere to this convention so that our files could be transported to another system for other analysis that we may not be capable of doing, or for sharing with other morphologists. Since the physical layout of the data is generic, but differ by

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619 455-9100

function (microscope data versus atlas data), our programs distinguish data files by reading in a section of signature code that points out these different functions. What all of this means to the user in the long run is that he does not have to concern himself about file maintenance or make notes on what every file pertains to. The computer system will keep track of this housekeeping for him automatically.

System Programs

1. Microscope System - these are a collection of programs that interface to the microscope camera and digitizer and provide a means of entering in anatomical information into the computer. This system is especially well suited for applications involving extensive three dimensional tracking of cellular bodies through the neuronal matrix. The data may be used to reconstruct the entire aspect of some physical structure as it courses through the brain, even through different Z sections. Coordinate information relative to a specified histochemical landmark is maintained in databases and all positional offsets are calculated to give absolute measurements during analysis.

2. Atlas System - these programs deal with the analysis of data as it pertains to the published atlas of various species. The two atlases currently available are for rat and monkey. The imaging system contains the model brain as outlined in these standard atlases. Any section of the brain may be recalled, rendering not only the morphological data but functional data as well. Templates may be created that contain references to these standard areas, but also permit the user to enter in morphometric data from the bitpad or microscope. The data may be as simple as an advanced level of notekeeping during routine perusal of tissue sections, or it may be applied to sophisticated methods of inter-species functional relationships, relevance to previously acquired data, both morphometric and textual (based on known literature), and three dimensional reconstruction and best fit into the standard models.

3. Data Analysis and Acquisition - Non-Digitized Image

Currently, data acquisition is operational from the live image, based on two techniques. One is the X,Y,Z coordinate representation of cells or cell types by symbols. There are over 2 million combinations programmed into the system, easily providing sufficient depth. The symbols themselves may have separate visual attributes, such as round, oval, polymorphic with major dendritic axes defined, extending the range to a magnitude on the order of 32 X 10^{12}. Every data record symbolized as such is easily treated to separate statistical or mathematical models. Positional and relational information is always available. Links to other data records provide information concerning projections to and from terminal fields, cell bodies, etc. In essence, this is a highly accurate notepad that keeps track of cellular locations. Double

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619455-9100

precision is used in the programs, giving 32 significant digits.

A second mode of non-digitized data acquisition already operational is based on vector images. The cell, or for that matter any structure in question, may be traced partially or entirely, from the microscopic image on the monitor regardless of the magnification being viewed. Vectors or lines are created in the computer graphics terminal that correlate with the movement and placement of the bitpad mouse. Streaming mode enters in vectors with daisy-chained endpoints, useful for very rapid tracing of outlines. Individual mode draws a separate vector for two distinct endpoints section. This mode is selected when this is non-contiguous areas need to be followed, or when following two different cell types sequentially. Nodal mode is a specialized case of streaming mode where the computer sets aside an array to maintain information of nodes as one is tracing the cell. When a node is reached, a option is selected to tell the computer that this is a branch point. The operator may continue tracing one of the two branches that bifurcate from the node, and later return to it for the other branch. This is extremely useful when traversing either the microscopic window (moving the microscope stage axes) or when going from slice to slice. Since the computer maintains absolute information on all coordinates, the operator is brought right back to the desired branch point. The node itself may have attributes assigned to it, so information can be readily retrieved concerning its history.

A background data menu is present at all times to assist the operator. Modes of data entry may be changed by moving the mouse to that window and pressing a mouse button. Similarly, the operator may request different screen attributes, such as decreasing the intensity of the camera video image, the background data menu, or the data planes themselves. Grids, rectangular boxes, or circles may be overlaid anywhere in the background to aid in sectoring off areas designated for analysis. All spatial relationships may be manipulated by the mouse as well, including translation, rotation, and scaling.

4. Other systems - Still in progress are the system levels that operate with the digititizer.

A) automatic densitometry of microscopic images - in applications of grain counting in autoradiography, the system can look at the 3 dimensional array that represents the image and make inferences based on the patterns. The simplest and fastest method is to count all data points that have their Z axis (or intensity level) above a certain threshold parameter. This would be counted as 'one' grain. The distribution of these grain counts can then be presented as a function of the XY space or as a ratio to 'non-grain' areas. This threshold parameter is highly variable, dependent on the quality of the optics, the method of preparation, and the quality of the method. The digitizer incorporates a Z axis threshold cutoff, whereby values on the video signal below this are equated

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619 455-9100

to zero. By noting a densitometric scale on the monitor, the user can adjust for this cutoff threshold and thus lower the background noise and improve the signal to noise ratio. The determination of what is grain and what is artifact is then carried out by the user, unfortunately contributing to high subjective bias and low reproducibility.

To mitigate these problems, the computer may be given the task of objectively distinguishing between grain and artifact. By analyzing the shape of the patterns, grains may be distinguished from artifact within a certain level of confidence. One method is to use shape algorithms that look at the contrast pattern of image boundaries and try to fit the above threshold data points within pre-defined patterns. Another method uses the intensity information and mathematically subtracts the edge boundary data point values until a single data point is isolated. That isolated point is considered the center of the grain and the count in the computer goes up by one. This method is useful for grains that take on an unusual shape. The patterns may be distinguished as single units, regardless of the shape.

2) automatic tracing of a metal impregnated cell body - by looking at the data points that cross the threshold parameter within a finite XY coordinate space, the contrast of the patterns can be determined. With this contrast information the Z axis can be followed, allowing the computer to automatically focus on different levels of the specimen. For this application, we need the Burleigh position translators. The entire cell projection may be followed and recorded in terms of boundary information under CPU control.

3) Image Enhancements - optical techniques exist for the enhancement of the microscopic image (polarization, phase contrast, dark field, interference contrast). They serve to improve the visualization of the image, especially boundary areas that delineate the pattern. Visual information before or after processing with these optical techniques may yet still yield useful information not available to the naked eye. The computer can take the image from the camera and process it further and redisplay the image in a new, enhanced form.

Algorithms that do video intensification, video enhanced contrast and edge enhancements, and pseudo coloring may be passed over the digitized image. Video intensification transformations sum the weaker images against the background noise. Since noise tends to be random, it can be averaged out from the signal that remains after each transformation. This is especially useful in real-time, live or dynamic situations where mottleness or muddiness can be readily removed. Video enhanced contrast and edge enhancement algorithms operate as described above in threshold detection, but will alter the data point intensity value around that threshold to render a more distinctive pattern. Pseudo coloring is used to give a higher degree of contrast definition not possible with gray scale images. Any gray scale level (defined by the data point intensity from 0 to

SCRIPPS CLINIC
AND RESEARCH FOUNDATION

10666 NORTH TORREY PINES ROAD
LA JOLLA, CALIFORNIA 92037
619455-9100

255) can be assigned a new color. That color may be an intense red, quickly bringing out visually the boundary information. Regions of different densities may be assigned very different colors to enhance their relationships with one another.

Another application during image enhancement is to render a population distribution of soma versus fibers in a single field, contingent on the ability of the histochemical stain to distiguish them. If the density difference between the two populations is sufficient to yield confidence, the computer can make the automatic analysis.

A GENERAL SYSTEM FOR COMPUTER BASED ACQUISITION, ANALYSIS AND DISPLAY OF MEDICAL IMAGE DATA

Daniel S. Schlusselberg
Wade K. Smith
Margaret H. Lewis
Bradley G. Culter
Donald J. Woodward

Departments of Cell Biology and Radiology
University of Texas Health Science Center at Dallas
Dallas, Texas 75235

ABSTRACT

A general computer-based system has been developed and implemented for acquiring and viewing medical image data. Originally developed for neuroanatomic studies (1), including investigations of cell topography and connectivity in brainstem nuclei, the system has become a versatile and powerful tool for three-dimensional analysis and display of a variety of types of image data, including studies of cardiac morphometry and 2-d gel electrophoresis. The system includes components for data input via video frame digitizer or digitizing tablet; graphical output through a high-resolution color graphics display or hardcopy plotter. Keys to the system's flexibility and power are a tree-structured data file system, in which line segments and shaded strips may be combined to form complex three-dimensional structures, and a disk-based virtual memory system which permits greater numerical accuracy and use of larger structures than would be otherwise possible with a 16-bit minicomputer.

Keywords and Phrases: Three-dimensional reconstruction, computer microscope, *graphics data base, shaded surface imaging, image analysis.*

I. SYSTEM DESCRIPTION

A. System components (Fig. 1)

The system has been assembled from standard components, based on a 16-bit minicomputer and a high performance graphics processor. The minicomputer (Data General S/130) has half a megabyte of main memory and a 200 megabyte disk; it is interfaced to a microprogrammable graphics processor (Ikonas RD3000) via a DMA link. These devices have been enhanced by our implementation of a disk-based virtual memory array system which permits manipulation of arrays too large to fit in locally addressable memory. Input to the Ikonas comes from a high resolution video camera (Cohu) on a 35mm film transport system (Mikel) or from an RCA video camera mounted on a light microscope. Output from the Ikonas goes to several video monitors by means of a Cohu Video Switching Matrix, which also controls the camera inputs to the Ikonas. These monitors include a Rantek high resolution color monitor which operates at 60hz, and a Conrac black and white monitor which operates at either 30 or 60hz. The Ikonas can display the contents of its frame buffer in two ways: as a 512 X 512 X 24 bit image in full color, or as a 1024 X 1024 X 8 bit image with 256 levels of gray or 4 pseudocolors. There are two Summagraphics BitPad graphics tablets for manual digitization with a hand held stylus. The tablet drives a user-defined Ikonas cursor for digitizing from video images which are stored in the frame buffer. Graphics output is also available from two Tektronix 4000-series terminals or from Televideo terminals with Tektronix emulation boards. Hardcopy output is available through a Versatec Matrix Printer/plotter.

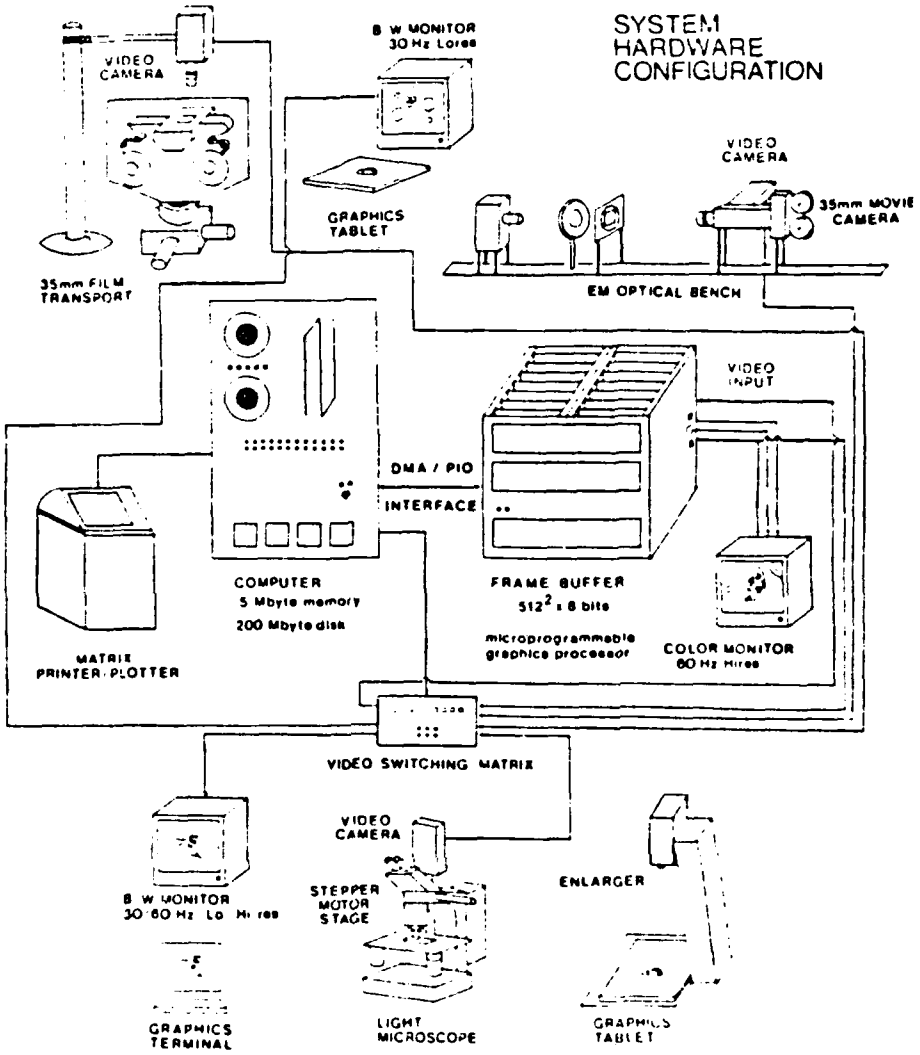


FIGURE 1.

Center: The host is a DATA GENERAL Eclipse S/130 16-bit minicomputer which is interfaced to an IKONAS RDS-3000 frame buffer with a high-speed 32-bit microprogrammable processor. VERSATEC matrix printer/plotter is the main high resolution hardcopy graphics output device. RAMTEK color monitor is the main graphics output device for the IKONAS frame buffer. COHU Video Switching Matrix manages routing of all video signals in laboratory.
Upper right: EM Optical Bench for aligning and copying serial electron micrographs into 35mm film strip sequences.
Upper left: MEKEL high-precision 35mm film transport which projects into a COHU high resolution video camera. CONRAC B/W monitor and SUMMAGRAPHICS digitizing tablet for display and data input.
Bottom: LEITZ Orthoplan light microscope with X/Y stepper motor stage and camera lucida drawing tube. RCA video camera linked to IKONAS video digitizer input. Graphics tablet, B/W monitor and TELEVIDEO terminal with PLOT-10 emulation board serve as graphics input and display devices. B/W monitor projects vector displays into light microscope viewing field through drawing tube.

B. Data acquisition

We are able to collect data from the light microscope, projecting microscope (enlarger), printed material, 35mm film strips, or mass storage video device. All programs are designed so that data from any of these sources can be entered at any time during the analysis of previously acquired data. Usually the data comes from biological objects which have been serially sectioned. The objects of interest may consist of anatomic boundaries, cell locations, autoradiographic grain positions or cell membranes and synapses in the case of serial electron micrographs. In the current system, object positions are entered interactively via a graphics tablet.

Before points are entered from any section a biological coordinate system is established which converts relative points from a graphics tablet into absolute points in the section. This involves calibrating the locations of tablet points in relation to the optic system which they pass through in order to appear on the video monitor. There are at least three coordinate systems which must be dealt with: 1) the tablet coordinate system, 2) the biological (absolute) coordinate system, and 3) the video monitor coordinate system (pixels). The biological coordinate system is defined by an origin and axis (usually Y-positive) and a scale in microns.

The light microscope data acquisition station (Fig. 1, bottom) operates as a "video lucida" system, analogous to the "camera lucida" system used to manually draw objects seen in the microscopic field. The drawing tube is aimed at a video display so that vectors which appear on the display are optically mixed with the image from the microscope slide. The user then sees a cursor and vectors overlaid on the image viewed through the eyepiece. The light microscope stage is translated in the horizontal plane by two computer-controlled stepper motors. In order to interpret the positions of objects seen through the microscope two additional coordinate systems are necessary: the stepper motor coordinate system and the light microscope coordinate sytem. The stepper motor coordinates are preset hardware locations which are defined in the interface to the computer. The light microscope coordinate system is invisible to the user but relates to the circular field that is seen through the eyepiece and any nonlinearities that may exist in the x-y plane. In addition, a third stepper motor controls the focus knob and provides z-coordinates for a biological specimen of sufficient thickness. A video camera mounted on the microscope introduces a camera coordinate system which is a subset of the microscope coordinate system.

In order to deal with all these coordinate systems, there are calibration and transformation subroutines to convert points in one coordinate system to another. For example, by using these routines a program can determine the number of stepper motor steps to move the image in the microscope to display a certain biological coordinate in the center of the microscopic field, or the biological coordinate of a pixel from a digitized video image of the microscopic field can be calculated. The calibration routines establish the matrices necessary for the transformation routines. Any time the stepper motors are moved, new matrices are generated to account for translations of all the coordinate systems. The user is able to monitor the calibration in the light microscope system because line segments and cell markers displayed on the video image should appear optically superimposed over the anatomic structures they represent in the microscopic field. This video image can be updated any time change in coordinate system occurs.

Similar programs are used to acquire data from the enlarger, 35mm film strip projector, video source, or directly from the graphics tablet. We use a specially designed optical light bench (2) to create 35mm film strips from serial electron micrograph negatives (Fig. 1, upper right). These film strips can then be examined with a Nokel Instruments precision transport system mounted on a platform with stepper motors controlling X-Y translation and rotation (Fig. 1, upper right). The computer also controls single frame movements of the filmstrip in forward and reverse directions. A Cohu high-resolution black and white camera is mounted above the projector and the images on the film strip are digitized to produce video images in the Ikonas frame buffer. These video images can be used to align serial sections using image combining techniques. Line segments can be traced over the video image by using a tablet controlled cursor, and stored in the file after user approval. When digitizing directly from the tablet (e.g. printed material laid on top of the tablet) the user traces over the drawing with the tablet's hand-held cursor and vectors appear on the video screen superimposed on a biological axis plotted on the screen. All data from the digitizing tablet, with appropriate calibration information, can be stored in data files for later analysis and display.

C. Three-dimensional database

We have developed a hierarchical file system for flexible interaction with data entered into the computer. All data is physically stored in a binary file as lists of four 32-bit floating point numbers which combine to form a logical tree structure. The virtual memory software package

can address any element in the file as if it were in a large 4-dimensional array. A 32-bit address is used to get the values of any 4-tuple in the data-base. The 32-bit numbers in the file can be grouped together to form 3-dimensional coordinates, address pointers to other elements in the file, matrices, and tree nodes. A node contains descriptive information and addresses (pointers) needed to traverse the tree structure. Each node contains the address of (pointer to) its parent node, as well as pointers to its siblings and first child (Fig. 2). Siblings are connected to each other as doubly linked lists, i.e. each node contains pointers to predecessors and successors. For example, to find the second sibling of a certain node one would have to get the address of the first sibling, and then get the address of that sibling's successor. There is also a pointer to a 4x4 transformation matrix which is associated with each node to allow for three-dimensional transformations (translation, rotation, scaling) at any level in the data-base.

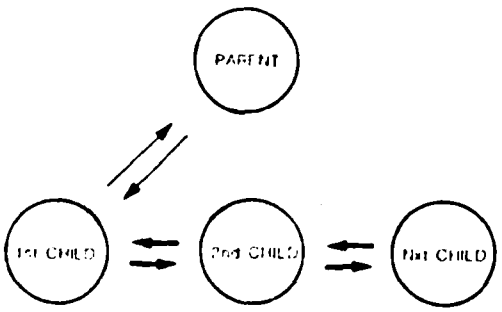


FIGURE 2.

Different levels in the hierarchy are represented by different types of nodes (Fig. 3). The first four entries of the file are reserved as a ROOT node, which contains global information about the file, and the addresses of the first SECTION and STRUCTURE nodes. Data acquisition programs must first create a SECTION node in order to begin collecting data for any given section. SECTION nodes contain information about the data acquisition hardware and biological material in the section in order to define a biological coordinate system. Data is entered for each section as SEGMENT nodes, which are the children of SECTION nodes. A segment is defined as any grouping of digitized points from material analyzed at any of the data entry stations. SEGMENT nodes point to lists of three-dimensional coordinates. The SEGMENT transformation matrix allows for local alignments to adjust for tissue distortion in the original material. Segments can include contours (line tracings around the border of a sectioned object), cell positions, autoradiographic grain positions, or fiducial marks (landmarks with relatively constant positions in serial sections which are used for alignment of serial sections).

Once segment and section data are entered, the next step in data analysis is to identify the segments from each section which form parts of structures. The segments are selected from line drawings on the video monitor by using a tablet controlled cursor. These selected segments become CONTOURs and are the children of SURFACE pointers. A set of contours can then be sent to a surfacing routine (3) which generates a set of triangles (STRIPS) which are also the children of the SURFACE pointer. This set of triangles, which connects contours from adjacent sections, defines the minimum surface area between the contours and generates input to a surface lighting routine (4). By combining the output of the lighting algorithm with perspective calculations, polygon fill, and Z-buffer hidden surface techniques we can construct a view of the three-dimensional surface of the object represented by a SURFACE node. An X-Y-Z axis and three-dimensional grids can be superimposed on the display to provide depth cues and scaling information. Each SURFACE node corresponds to one continuous branch of a structure, i.e. for each BRANCH node there is one SURFACE node. If a structure does not branch its STRUCTURE node will only point to one BRANCH node. On the other hand, a branching structure will be represented by a hierarchy of BRANCH nodes for each branch.

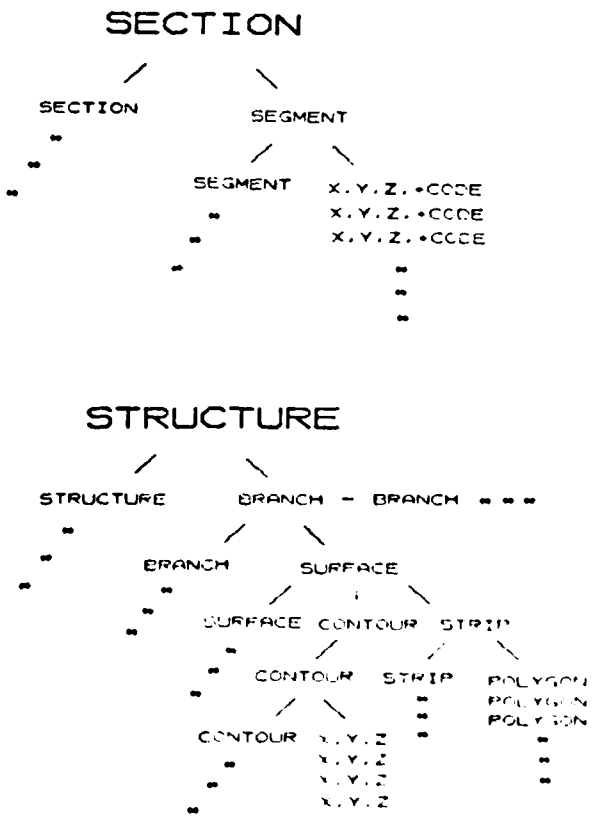


FIGURE 3.

A STRUCTURE does not necessarily have to represent a surface, since its contours can be formed from any type of SEGMENTS. Groups of cells or autoradiographic dots can also be organized into structures for further analysis. The programs include quantitative algorithms for counting objects in space or calculating volumes and surface areas. Input to these algorithms can come from various node types.

D. Display

There are several ways we can display these collections of data. Hardcopy is available through a Versatec Matrix Printer/plotter as text or vectors. Depth in three-dimensional views is simulated by having different line thicknesses for vectors that represent different distances to the viewing eye (Fig. 4) that represent different distances to the viewing eye. Vector output is also available on Tektronix 4014 or 4012 terminals for analyzing individual sections. For color and shaded graphics, the Ikonas display system is used. The Ikonas has low and high resolution modes of 512 X 512 X 32 bit and 1024 X 1024 X 8 bit pixels respectively. In high resolution mode the Z-buffer is stored in the lower half of the image, giving 1024 X 512 pixel resolution images with hidden lines and surfaces removed. In programs which have three-dimensional graphics capability a particular view of objects in biological space is defined by eye, viewpoint, and light vector coordinates. Clipping planes and a lens setting can be used to modify the view by selecting out different parts or zooming in and out.

The Ikonas graphics system contains a microprogrammable processor with highly parallel operation for rapid image construction and analysis. Microprograms have been implemented for rapid vector generation and smooth-shaded polygon filling. These programs speed up the time it takes to construct three-dimensional views of complex anatomical structures, and help the user select ideal views interactively. The Ikonas uses a 32-bit color lookup table to enable pseudocolor or full-color operation. This lookup table can also be used to set thresholds for edge detection or grain counting algorithms.

II. DISCUSSION

The system has been used by several laboratories in the medical school, with varied applications. One project involves reconstructing the locations of pigmented dopamine neurons in human brainstems, and counting regional cell numbers in different parts of the brain (Fig. 5) Our computer-based system was necessary because of the large number of dopamine containing neurons in the human brain, and the necessity to analyze the number of these neurons in different locations with respect to aging and neurological disease. A similar project involves the display of neurons in the ventral thalamic nucleus that project their nerve endings into the somatosensory cortex of the rat (Fig. 6). The system was used to assess the three-dimensional topography present in groups of cells which project to similar areas. Three-dimensional surface display techniques have been used to reconstruct neuroanatomic fiber tracts and cell groups in the human brain for

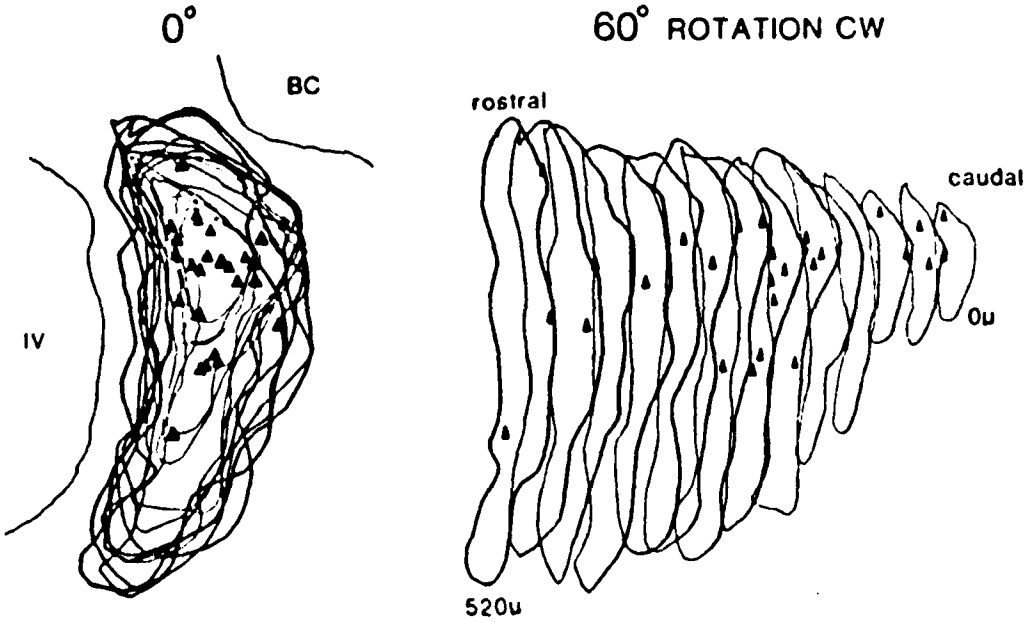


Figure 4: Example of hardcopy output from Versatec plotter. Vectors represent outlines of anatomic boundaries with different line thicknesses to indicate depth. Filled triangles are specific cells which have been identified under the microscope.

teaching purposes (Fig. 7). We have also been digitizing serial sections of canine hearts with myocardial infarcts in order to compare the reconstructed images and volume estimations with those obtained from nuclear scans (Fig. 8). Another application has been quantitation of protein distributions in 2-d gel electrophoresis images. The gel is photographed on 35mm film, and individual frames are digitized into the Ikonas frame buffer by a video camera mounted on the film transport. A special program can then detect spots of accumulated protein in the digitized image and report protein presence and amount in the sample.

Because of our flexible filing system and modular programming under a time-sharing system, the programs are easily adapted for many applications. Although we have incorporated many design strategies used in other systems (5,6,7), we have created our own unique data base with identical I/O management subroutines in all programs. These I/O routines allow programs to access disk files as if they were large arrays, and by using a 32-bit address we can vastly exceed the host memory limits. The shared-page facility allows disk buffers that are outside of program address limits, but still within the host memory, to be remapped into program address space. This facilitates acceptable program responsiveness and speed. We have taken advantage of the rapid display capability of the Ikonas graphics system while retaining the flexibility of its microprogrammable processor. Our approach to three-dimensional reconstruction and analysis has been to create high-resolution images with high information content as depth cues. Most real-time display systems impose limits on the complexity of data they can handle and the information they provide is not maximally expressed in a photograph of a single frame. They are useful, however, in previewing three-dimensional views of objects and in establishing the view parameters for other programs. By creating images with smooth shaded surfaces and utilizing hidden surface techniques instead of dynamic vector displays, we can produce a three-dimensional display in a publishable form.

Currently, all programming is done in Fortran V under the AOS Operating System, with some low-level driver routines written in Eclipse assembly language. We are in the process of re-writing the data acquisition, database management, and user interface routines in the C programming language; and are considering implementation of the CORE system graphic standards as outlined in the quarterly report of Siggraph-ACM (8).

Future plans include implementation of general purpose algorithms for computer detection of cell boundaries (9), grain counting and electron micrograph analysis. Microprogrammed versions of these algorithms could run quickly enough for user interaction in a semi-automatic system. The light microscope system is also being equipped to do three-dimensional analysis of dendrite branching patterns in Golgi stained neurons.

By implementing a general and flexible system which can interface to a variety of special instruments, we have thus developed a powerful tool for computer analysis of biological material at both microscopic and macroscopic levels.

Acknowledgements:

This work was supported by Biologic Humanics Foundation and the National Science Foundation (NSF BNS77-01174).

REFERENCES

1. WK Smith, DS Schlusselberg, and DJ Woodward. A Computer System for Neuroanatomical Data Acquisition, Analysis, and Display. Society for Neuroscience Abstract 135.18, October, 1981.

2. JK Stevens, TL Davis, N Friedman, P Sterling. A Systematic Approach to Reconstructing Microcircuitry by Electron Microscopy of Serial Sections. Brain Research Reviews, 2 (1980) 265-293.

3. H Fuchs, ZM Kedem, S Uselton. Optimal Surface Reconstruction from contours. Comm ACM 20:693-702, 1977.

3. JF Blinn. Models of Light Reflection for Computer Synthesized Pictures. Computer Graphics, 11(2) 1977.

5. C Levinthal, E Michaud, and C. Tsouras. Computer-aided reconstruction from serial sections. Fed. Proc. 33(12) 2336-2340, 1974.

6. RD Lindsay. Computer Analysis of Neuronal Structures, in Computers in Biology and Medicine, Plenum Press, 1977.

7. TA Woolsey, MI Dierker. Computer-assisted Recording of Neuroanatomical Data, in Neuroanatomical Research Techniques, Academic Press, 1978.

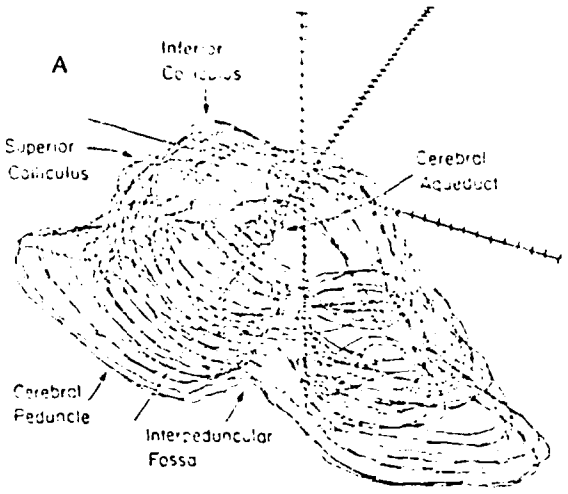
8. Status Report of the Graphic Standards Planning Commitee. Computer Graphics, 13(3) 1979.

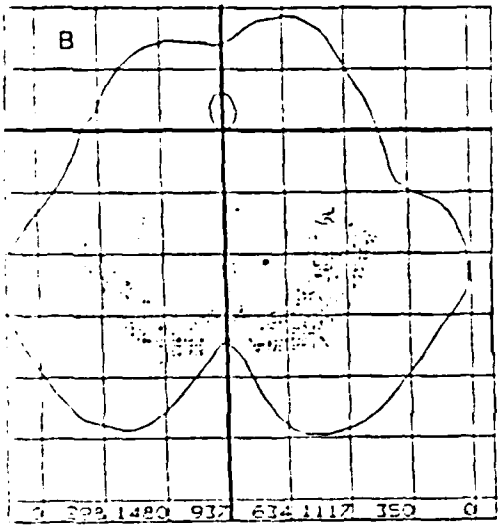
9. DS Schlusselberg, WK Smith, B Culter, DJ Woodward. A Computer System for Semi-automatic Cell Recognition in Neuroanatomic Studies. To be presented at Neuroscience Society Annual Meeting, October 1982.

10. MH Lewis, DS Schlusselberg, WK Smith, HK Hesler, DJ Woodward, and HI Siegel. Three-Dimensional Cardiac Morphometry with Computer Graphics. To be presented at Computers in Cardiology, 1982.

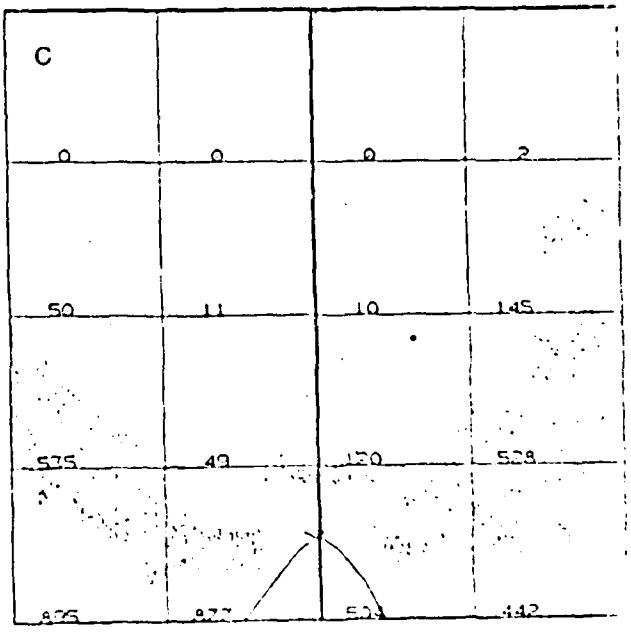
Figure 5: Computer analysis and reconstruction of dopamine neuron distributions in a human brainstem.

A. Outlines of gross brain structures were entered into the computer by using the projection microscope. Objects of interest are traced manually on the graphics tablet after entering origin and axis to define biological coordinate system in microns. Serial sections are then aligned by visual inspection, and can be plotted by the three-dimensional graphics package. Axis scale is 1000 microns, no depth shading. Each section was 200 microns thick.





B. Computer analysis of a single section in which the locations of all pigmented dopamine containing cells have been entered through the light microscope. The data acquisition program uses stepper motors to scan a user-defined area. Cell locations which have been identified are displayed in the microscopic field by aiming the drawing tube at a video monitor graphics system which has been calibrated to the biological coordinate system. This allows the user to see in the microscope an overlaid image which indicates which cells have been selected. When all cell locations have been entered the user can get a summary diagram as shown here. A grid has been drawn at 4000 micron intervals, and the computer has been instructed to count all the cells in each column. Cell locations are represented by a single dot.



C. By creating a smaller biological window to view, the image is blown up to show finer detail. The 400 micron grid is once again plotted, and the computer instructed to count the number of cells in each grid box.

This data is part of a study to determine the absolute number of dopamine cells in specific regions of the brain, and was provided by Dr. Dwight German, Department of Psychiatry.

Figure 6: Display of cells in the ventral thalamic nucleus of the rat which project to a specific region in cortex. Cell locations are indicated by spheres with a diameter approximating that of the cell body and its branches. The lines represent the boundaries of anatomic regions which were traced in the microscope. Data provided by Dr. John Chapin, Department of Cell Biology.

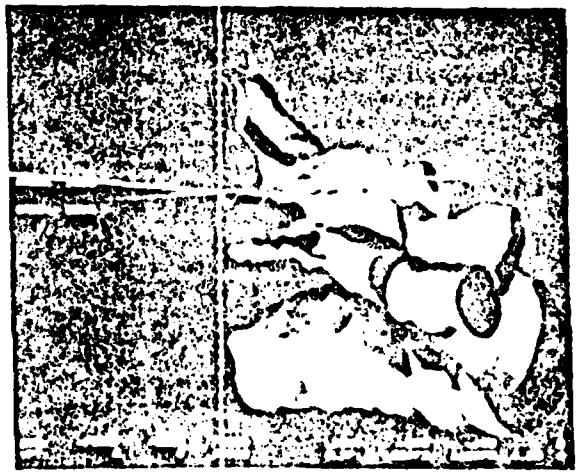


Figure 7: Three-dimensional reconstruction of fiber tracts and cell groups in the human brainstem, derived from a stereotaxic atlas of the human brain. An example of smooth-shaded surfacing techniques with hidden surface removal. Axis scale is 1000 microns. Three-dimensional viewing parameters are indicated at bottom of video screen.

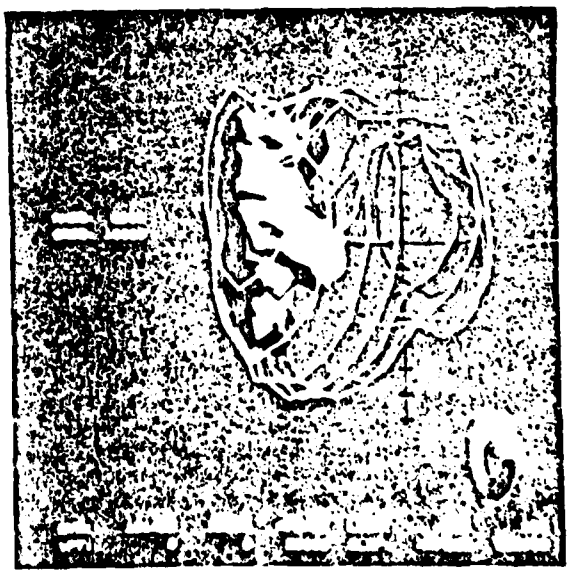


Figure 8: Reconstruction of a myocardial infarct experimentally induced in a dog (10). Posterior view of heart with infarct in left ventricle on the left. This infarct had large anterior and lateral branches and a small posterior branch, seen closest to the viewer. Vectors represent exterior outlines of the heart, with depth shading. At lower right is the colormap used to generate the black and white image. This data comes from a 35 mm film of serial sections taken through the heart and specifically stained for infarcted tissue.

Reprinted from

PROCEEDINGS OF THE FOURTH ANNUAL CONFERENCE AND EXPOSITION OF

THE NATIONAL COMPUTER GRAPHICS ASSOCIATION

McCormick Place
Chicago, Illinois
June 26-30, 1983

Hierarchical Database Design for Biological Modeling

Dr. Wade K. Smith
Dr. Daniel S. Schlusselberg
Mr. Bradley G. Culter
Dr. Donald J. Woodward

Department of Cell Biology
University of Texas
Health Science Center at Dallas
5323 Harry Hines Blvd.
Dallas, Texas 75235

Dr. Eric R. Lacy

Department of Anatomy
Harvard Medical School
25 Shattuck St.
Boston, Mass. 02115

ABSTRACT

Ongoing studies in this laboratory have explored the application of computer graphics techniques for 3-dimensional reconstruction of biological structures from serial slices (1,2,3,4). The design of a database for storage, manipulation and display of the serial section data has been the dominant task, while software for graphical input-output has been straightforward. We describe here the evolution of our database design strategies to what we now consider a highly flexible system for managing our biological modeling data. The results of one task, imaging of a complex multiple tubular system in a kidney, illustrates the power of the database.

INTRODUCTION

There is a growing interest in the use of computer technology for the digitization of biological material, and for three-dimensional reconstruction of anatomical objects. Several laboratories have developed computer systems for these tasks (9,10,11), but often they have concentrated on limited applications. These systems have contributed to the evolution of data acquisition and display strategies, some of which have been incorporated into our system. All of these systems have a form of database management for the storage and retrieval of biological coordinate information. After surveying these methods we felt a need to develop an advanced general-purpose database for the storage of anatomical information. Our long-term goal is to promote the evolution of database management of biological information, so that it may become more useful for the scientific community.

DATA ACQUISTION

We are able to collect data from the light microscope, projecting microscope (enlarger), printed material, 35mm film strips, or mass storage video device. All programs are designed so that data from any of these sources can be entered at any time during the analysis of previously acquired data. The objects of interest may consist of anatomical boundaries, cell locations, autoradiographic grain positions or cell membranes and organelles in the case of serial electron micrographs. In the current system, all object positions

and boundaries are entered interactively via a graphics tablet and video cursor.

Before points are entered from any section an anatomical coordinate system is established which converts relative points from a graphics tablet into absolute points in the section. This involves calibrating the locations of tablet points in relation to the optic system which they pass through in order to be superimposed over a tissue section (Fig. 1). There are at least three coordinate systems which must be dealt with: (1) the tablet coordinate system, (2) the anatomical coordinate system and (3) the video monitor coordinate system (pixels). Transformation subroutines convert points from one coordinate system to another. Calibration subroutines determine the matrices necessary for the transformation

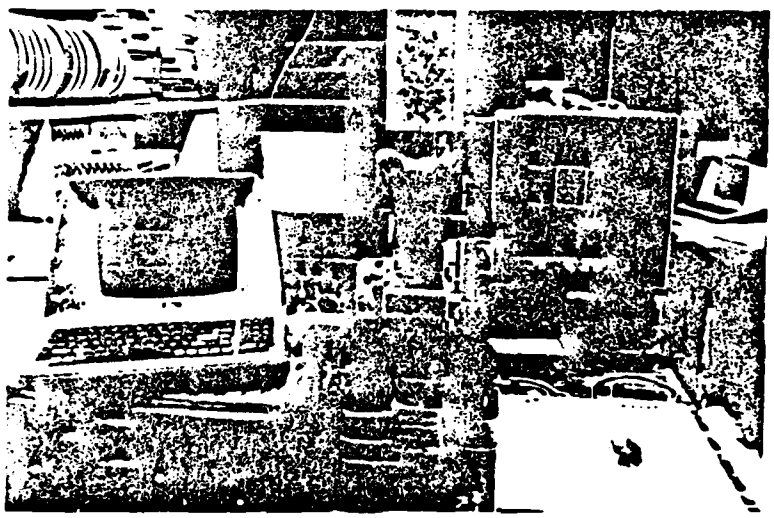


Figure 1. Example of a digitizing station.

Illustration of the video lucida computer microscope station. The camera lucida drawing tube is aimed at a vector graphics display on a video monitor. In this way, lines which appear on the video screen are also seen when viewing a microscope field through the eyepiece. After calibration of a biological coordinate system on the microscope slide, the computer plots lines which optically superimpose boundaries in the tissue specimens. Cell positions are indicated by triangles. Computer controlled stepper motors allow for movement of the specimen within the microscopic field and updating of the overlying graphical images through the drawing tube. A video camera is mounted above the microscope for digitization into a frame buffer. The digitized image is seen on the video monitor on the left.

The hardware includes a LEITZ Orthoplan Microscope, XY stepper motor stage, stepper motor focus control, graphics terminal and tablet, 30Hz video monitor for video lucida graphical display, 60Hz video monitor for frame buffer display, DATA GENERAL S/130 16-bit minicomputer, and IKONAS RD-3000 color graphics system with frame buffer and 64-bit microprogrammable bit-slice processor.

routines, and generate a set of paramaters which can be stored in the file. These parameters are used to assist in recalibrating an input device so that vector displays of anatomical information in the file can be optically superimposed over the biological material they represent. This allows confirmation of accuracy and entry of new data at any time. When programmed changes occur in locations of physical objects within the digitizing system, the coordinate transformation matrices must be appropriately modified.

Often the need exists to enter data from the same physical material using different input devices. For example, the gross anatomic boundaries in a section of human brain are too large for the light microscope, and must be traced by projecting an image of the section onto a graphics tablet using a photographic enlarger (projecting microscope). In order to then digitize cell positions within the same sections, the data must be reexamined under the light microscope with the preservation of an anatomical coordinate system common to both input devices. This involves recalibration using digitized points already stored in the data file, and finding their corresponding physical locations in the current input device.

DATA STORAGE

Two-Dimensional Models of Anatomic Structures

The most common two-dimensional representation of an anatomical object is a line traced around its external boundary as viewed in a serial section. We store these line drawings as ordered vertex lists, where the first point in the list is graphically represented as a "move" (pen up), and subsequent points are "draws" (pen down). Similarly, graphical symbols are used to represent single point data such as cell positions, autoradiographic grains, synapses, etc. Symbol definitions are normalized coordinates stored in the program or data file and are translated to an anatomical coordinate position for final display. A filled region, such as a pigment containing cell body, can be represented by center of mass, area, and intensity, when using digitized video images for data acquisition. Digitized information from one particular section is organized into "segments". Segments include lines (vertex lists), object locations (lists of single points) and fiducial marks for serial section alignment.

Computer Display Techniques

The graphical display of digitized information requires the definition of an anatomical window through which a part of the anatomical coordinate area is viewed. This requires clipping out portions of data that are outside of the window. The anatomical window is then mapped to a physical display device by using display coordinates which define a viewport. The window can be combined with images of objects seen in data input devices or manipulated by the user to focus in on certain regions and create summary diagrams of all data entered. The current parameters used to define this window are stored in the data file, so that it is not necessary for a user to redefine them each time a data file is examined.

Three-dimensional Models of Anatomical Structures

There are several strategies for creating three-dimensional models of previously digitized anatomical data. By specifying certain viewing parameters and applying a perspective transformation (5), display coordinates are generated from vertex lists to create three-dimensional line drawings. Depth perception in these displays can be enhanced by using line intensity or thickness to represent distance from the eye or by hidden-line algorithms, which assume that each vertex list represents an opaque polygon.

Recently, new techniques have evolved for computer reconstruction of three-dimensional surfaces of objects which have been serially sectioned. Triangulation algorithms are used to generate an ordered polygon list from two vertex lists representing two "contours" of an object. A contour represents the external boundary of an object which has been sectioned by a plane. Contours are derived from segments in sequential sections. There are several requirements of triangulation algorithms which must be satisfied by the database. For any given section, several segments can be digitized, representing different objects that have been sliced in that section. Each segment that belongs to one object will generate one contour. Techniques are necessary to specify which contours belong to the same object in different sections. If an object branches, it will generate more than one contour in each section. Most triangulation algorithms will not generate a proper set of polygons for branching surfaces when given more than two contours as input. It is also necessary to keep track of the branching hierarchy in biological objects, since this may contain important structural and functional information. The simplest way to accomplish this is to store a set of sequential contours as one branch, which represents one three-dimensional surface. Each branch serves as input to a triangulation algorithm, since it only contains one contour per section.

When an object branches, it is necessary to construct a hierarchy in the database to represent the relationship between branches. If one branch splits into two branches, the three branches will share a contour in one of the sections (Fig. 2). To satisfy the triangulation requirements this shared contour must be present in all three

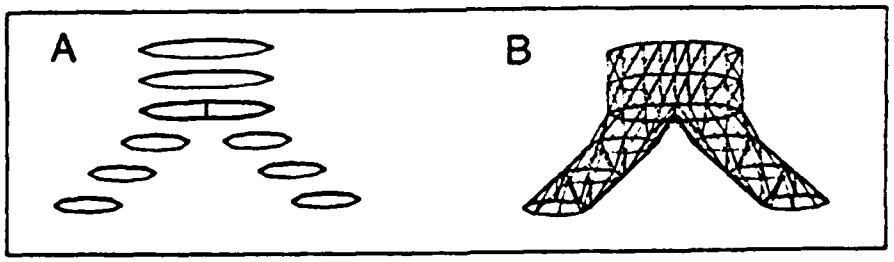


Figure 2. Example of a branching object.

A) A branching structure has been cut in serial sections and its boundaries (contours) are depicted in the figure. Notice that the third contour from the top is shared by all three branches.
B) In order to triangulate the surface of the branches properly, the shared contour must be split with a line segment into three contours: (1) the original complete contour for generation of polygons in the lower strip of the upper branch, (2) and (3) the two new split contours which are used to generate polygons for the lower branches.

branch definitions, but should be split in two of the branches. Various techniques for automatic and semi-automatic splitting of contours for branches exist. In our current implementation we rely on the user to specify where to split the contour, because of the unpredictability of contour complexity in serially sectioned biological material. The triangulation algorithm also requires that a contour should consist of one continuous vertex list, so that different segments which are part of the same contour must be appended together.

The polygons generated by the triangulation algorithm are stored along with their normal vectors, in order to use lighting and polygon fill algorithms to create three-dimensional surfaces in a raster display system. Each ordered set of polygons generated between two contours is called a "strip". A branch, therefore, contains an ordered set of contours with a corresponding ordered set of polygon strips. For continuous smooth shading of a polygon, average normals for each vertex in the polygon must be calculated. Vertices are assigned average normals based on the normal vectors of all polygons which share the vertex. This is done only once, when the entire polygon representation of an object's surface has been constructed. The program which calculates average normals uses the hierarchical tree structure of branch definitions to determine which strips in a branch are likely to have polygons which share vertices, thus eliminating the need to scan every vertex in the complete polygon definition of an object surface. Both planar normals and average normals should be stored in the database, to allow the option of displaying either polygon shading or smooth Gouraurd shading (6).

DATABASE MANAGEMENT: CURRENT IMPLEMENTATION

Our current programming environment is in the 'C' programming language (13), under the AOS Operating System, for a Data General Eclipse S/130 16-bit minicomputer with a 200 megabyte disk and magnetic tape backup. We have developed a virtual file system for flexible interaction with data entered into the computer; this allows programs and data to vastly exceed host memory limits and simplifies I/O management. All data is physically stored in a binary file which is treated as an extension of the program's address space. The virtual memory software package can address any element of the file through 32-bit pointers to byte locations. Data records are organized into nodes which contain information about the type of data stored and address pointers to lists of graphical data and other nodes. The structure of the graphical data list depends upon its type; for example, a segment consisting of a line points to a list of data triplets. Each triplet uses 32-bit floating point numbers to represent X, Y and Z coordinates in 3-space.

Our hierarchical node structure employs the Knuth transform of an n-ary tree (7). Each node contains the address of (pointer to) its parent node, as well as pointers to its siblings and first, last and current children (Fig. 3). However, siblings are connected to each other as doubly linked lists, i.e. each node contains pointers to predecessors and successors. To find the second sibling of a certain node one would have to get the address of the first sibling, and then get the address of that sibling's successor. The current child is used as the default for certain node operations. The presence of the current and last children also serves to speed node traversal times through shortest path computations.

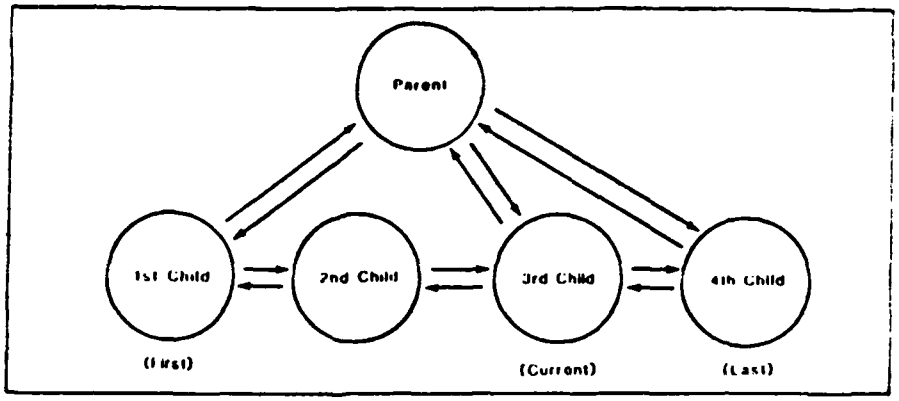


Figure 3. Links used to create a hierarchical tree structure.

Hierarchical trees with multiple branches are created by setting aside links for parent-child relationships, and for predecessor-successor (sibling) relationships. All nodes point to their parents, and parents point to first, current and last children. Siblings are all doubly linked.

Different levels in the hierarchy are represented by different types of nodes (Fig. 4). The first four bytes in the file contain a pointer to the first free byte in the database. Immediately following this end-of-file pointer is the ROOT node, which contains global information about the file and pointers to a group of FAMILY nodes. Included among the FAMILY nodes are SECTIONs and OBJECTs. Data acquisition programs must first create SECTION nodes for each section to be analyzed. Data is entered for each section as SEGMENT nodes, which are the children of SECTION nodes. A segment is defined as any grouping of digitized points from material analyzed at any of the data entry stations. SEGMENT nodes point to lists of three-dimensional coordinates. The SEGMENT transformation matrix allows for local alignments to adjust for tissue distortion in the original material. Segments can include lines (e.g., tracings around the border of a sectioned object), cell positions, autoradiographic grain positions, or fiduciary marks (landmarks with relatively constant positions in serial sections which are used for alignment).

Once segment and section data are entered, the next step in data analysis is to identify the segments from each section which form parts of biological objects. The biological object is represented by an OBJECT node, which points to the first BRANCH node. If a structure does not branch its OBJECT node will only point to one BRANCH node. On the other hand, a branching structure will be represented by a hierarchy of BRANCH nodes. Each BRANCH node points to a family of STRIP nodes, which represent the surface definition of that branch. Selected line segments in each section become CONTOURs and are grouped in pairs for each STRIP node. Each pair of contours belonging to a STRIP node can then be sent to a triangulation routine which generates an ordered set of polygons called a strip. The strip of polygons, which connects contours from adjacent sections, defines the minimum surface area between the contours (8).

There are two types of surfaces which can be generated by the triangulation algorithm, as specified in the STRIP node: (1) an open surface in which the ends of the surface do not meet, as in a sheet, (2) a closed surface in which the ends do meet, as in a tube.

There are two other families of the ROOT to note here. One is the ORPHAN family. As one builds an object from a series of digitized tissue slices, branches are often found before it is known where they are to be placed in the branch hierarchy. In our earlier strategies, one had to immediately attach new branches to their parents. By keeping isolated branches in the orphan family, users can now build objects in fewer passes through the serial sections. The other family to note is the GHOST family. As nodes are deleted from the tree, they are placed in this repository so that users might have a chance to return them to the hierarchy. A repacking, or garbage collection, operation clears the ghost family and removes the node and its data from the file along with all its descendents' nodes and data.

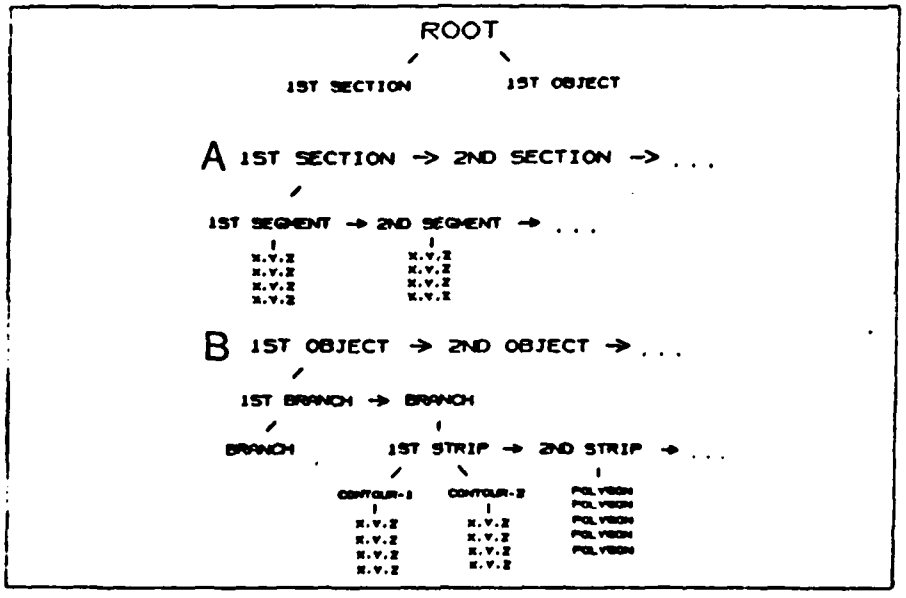


Figure 4. Hierarchy in 3-dimensional anatomic database.

A) SECTION nodes are considered siblings, and are doubly linked. SEGMENT nodes are the children of SECTION nodes, and also contain pointers to lists of three-dimensional coordinates in the data file.
B) An OBJECT node represents the complete surface description of a three-dimensional structure. OBJECT nodes are connected as siblings; their children are BRANCH nodes. BRANCH nodes connect to each other in a hierarchical fashion, depending on the branching nature of the structure. Each BRANCH node points to a family of STRIP nodes, which contain the surface description of that BRANCH. The STRIP nodes point to pairs of CONTOUR nodes and to ordered polygon lists, which are groupings of three-dimensional coordinates for vertices and normal vectors. CONTOUR nodes point to three-dimensional coordinate lists in a similar manner as SEGMENT nodes.

Commands exist for the manipulation of data in the hierarchical tree. Users can split, append or delete line segments that are already stored in the file. Capabilities also exist for removal of excess points and for the inversion of the order of points in a vertex list. Segments are identified by number, label or by placing a video cursor near their centers after displaying them within a two-dimensional biological window. A file editor exists for direct modification and display of the real numbers in the file. Elements in the hierarchical file can be plotted in either two-dimensional or three-dimensional forms (Fig. 6). The plotting subroutines will expand all descendents of the desired node. Quantitative commands work in a similar manner: nodes and descendents are used as input.

```
structure {
integer TYPE;                /* Type of node (e.g. ROOT, SECTION,
                                SEGMENT, OBJECT, etc.) */
integer SUBTYPE;             /* Node subtype (e.g. What kind of
                                SEGMENT is this?) */
filepointer LABEL;           /* Address in the file of a character
                                string identifying this node */

nodepointer PARENT;          /* Pointer to the node which is the
                                parent of this node in the hier-
                                archical tree (Pointer = Address
                                of beginning of node in file) */
nodepointer PREVIOUS;        /* Pointer to the previous sibling */
nodepointer NEXT;            /* Pointer to next sibling in list */
integer SONS;                /* Number of offspring of this node */
nodepointer FIRST_SON;       /* Pointer to first offspring */
nodepointer LAST_SON;        /* Pointer to last offspring */
nodepointer CURRENT_SON;     /* Pointer to current offspring */

integer RECORDS;             /* Number of data records that belong
                                to this node */
integer RECORD_TYPE;         /* Type of data record (used to select
                                a template with which to read data,
                                and specifies record_size) */
filepointer FIRST_RECORD;    /* Address in file of first record
                                (Address of next record is:
                                FIRST_RECORD + RECORD_SIZE) */

filepointer MATRIX;          /* Address of a 4 X 4 transformation
                                matrix */
float RED, GREEN, BLUE;      /* Color of graphical data that is
                                pointed to by this node */
integer FLAGS;               /* 16 individual bits are available
                                to turn various features on and off
                                for any given node */
integer VARIABLES[21];       /* Each node can have a set of variables
                                or room for future expansion */
} VIRTUAL_NODE;
```

Figure 5. Data structures used to create nodes in virtual file.

The C programming language allows structured data formats, which we have utilized in the creation of a virtual node description. All nodes in the hierarchical tree, including the ROOT node, use this format.

DISCUSSION

There is a paucity of literature available concerning database management for computer digitization of anatomical structures. To our knowledge, this work represents the first effort toward the development of an advanced general-purpose database system for this nascent field. It appears that this kind of graphics database is similar in many ways to that required for CAD/CAM systems. In particular, we have found that many features of the IGES standards for transfer of CAD information overlap those necessary for storage of structural data on biological objects. As these various systems mature, we expect that a more generalized set of requirements for graphical database design will develop.

Large amounts of disk space and memory are clearly desirable. Single data files to date have occupied up to 10 megabytes of memory on disk. A large disk is clearly required, since thousands of vertices are needed to portray accurately three-dimensional images of biological objects. Most of this space is dependent on the detail required of the line drawings (vertex lists) that have been traced around objects of interest. Simple circular boundaries, often seen in serial electron micrographs, can be entered by manually selecting individual points with the tablet and visual cursor; while complex boundaries, such as the external surface of human cortex, require a stream of points to be generated while the cursor traces the outline. We have already accumulated data files with contours containing up to 500 points, and sections containing up to 100 contours. The triangulation algorithm has the potential for increasing the data file size by a factor of nine, because of the space required by polygon definitions of surfaces.

We used this system in a recent task to reconstruct a complex multiple tubular system in the kidney of the elasmobranch little skate, Raja erinacea (14,15). Manual digitization of ~3100 cross-sectional profiles of the tortuous system of kidney tubules, observed in a series of 125 consecutive slices, required three days of work, while another two days were spent building the object definition of this complex looping structure. The principal anatomic feature of this system is a bundle of tubules which are arranged in a tight parallel array. These tubules do not travel a straight path, and instead change directions repeatedly. Such a tubular system is considered in the database as a tree-like structure with single portions connected by nodes. New nodes are defined each time the tubule changes its direction. A tubule is, in effect, a path through a tree which does not branch. An image of the final reconstruction shows the bundle of tubules coursing its path within a connective tissue sheath (Fig. 6).

To summarize, we have implemented a virtual memory software package combined with C-based data structures to a hierarchically organized three-dimensional anatomic database. This database serves data acquisition and analysis programs, as well as three-dimensional reconstruction and smooth shaded graphical display programs.

ACKNOWLEDGEMENTS

This work was supported by the Biological Humanics Foundation to DJW and the NIAAA (1F32AA05198-01) to DSS.

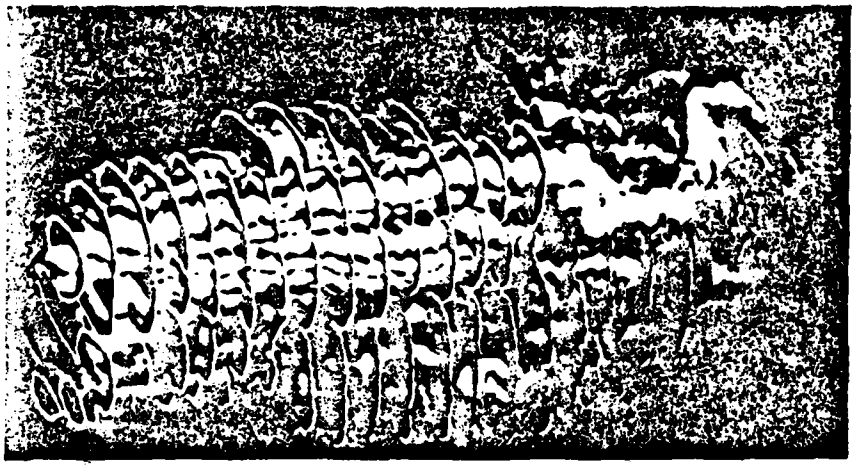


Figure 6. Reconstruction of portion of nephron from the kidney of the elasmobranch little skate, Raja erinacea.

Data consists of approximately 100,000 polygons generated from 3100 tubule profiles through 125 serial slices. This portion of the nephron is a specialized bundle of five tubules arranged in a parallel array. This bundle is thought to be responsible for the ability of the skate kidney to regulate body fluids in sea water.

REFERENCES

1. WK Smith, DS Schlusselberg, and DJ Woodward. A Computer System for Neuroanatomical Data Acquisition, Analysis, and Display. Society for Neuroscience Abstract 135.10, October 1981.

2. DS Schlusselberg, WK Smith, MH Lewis, B Culter, and DJ Woodward. A General System for Computer-Based Acquisition, Analysis and Display of Medical Image Data. ACM Conference Proceedings, ACM Annual Meeting, October 1982.

3. DS Schlusselberg, WK Smith, B Culter, DJ Woodward. A Computer System for Semi-automatic Cell Recognition in Neuroanatomic Studies. Society for Neuroscience Abstract 182.9, October 1982.

4. DC German, DS Schlusselberg, BA McMillen, K McDermott, WK Smith, DJ Woodward . Asymmetries in Human Brain Dopamine Receptor Binding: Relationship to 3-Dimensional Reconstruction of Midbrain Dopamine Neurons. Society for Neuroscience Abstract 30.3, October 1982.

5. WM Newman, RF Sproull. Principles of Interactive Computer Graphics. McGraw Hill, 1973.

6. H Gourand. Computer Display of Curved Surfaces. IEEE Trans Comput (20):623-629, 1971.

7. JL Pfaltz. Computer Data Structures. McGraw Hill, 1977.

8. H Fuchs, ZM Kedem, S Uselton. Optimal Surface Reconstruction from Contours. Commun ACM 20:693-702, 1977.

9. JK Stevens, TL Davis, N Friedman, P Sterling. A Systematic Approach to Reconstructing Microcircuitry by Electron Microscopy of Serial Sections. Brain Res Rev (2):265-293, 1980.

10. R Llinas, DE Hillman. A Multipurpose Tri-dimensional Reconstruction Computer System for Neuroanatomy. Golgi Centennial Symp Proc (M Santini, ed.) Raven Press, NY 1975.

11. ER Macagno, C Levinthal, I Sobel. Three-dimensional Computer Reconstruction of Neuronal Assemblies. Ann Rev Biophys Bioengng, (8):323-351, 1979.

12. MH Lewis, DS Schlusselberg, WK Smith, HK Hagler, DJ Woodward, LM Buja. Three-Dimensional Cardiac Morphometry with Computer Graphics. Computers in Cardiology Conference Proc, October 1982.

13. BW Kernighan, DM Richie. The C Programming Language. Prentice-Hall, 1978.

14. ER Lacy, B Schmidt-Nielsen, RG Galaske, H Stolte. Configuration of the skate (Raja erinacea) nephron and ultrastructure of two segments of the proximal tubule. Bull. Mt. Desert Is. Biol. Lab., (15):56-58, 1975.

15. H Stolte, RG Galaske, GM Eisenbach, C Lechene, B Schmidt-Nielsen, JW Boylan. Renal Tubule Ion Transport and Collecting Duct Function in the Elasmobranch Little Skate, Raja erinacea. J. Exp. Zool., (199):403-410, 1977.

3. Brain Image Modeling and Morphometrics

Computer Modeling in Radiology and the Anatomical Sciences

D. J. WOODWARD, W. K. SMITH and D. S. SCHLUSSELBERG

Department of Cell Biology
University of Texas Health Science Center at Dallas
Dallas, Texas 75235 U.S.A.

Summary

The Anatomical Sciences in general, in parallel with Radiology, have long recognized the need for computer-assisted technology. Initial computer applications began in the late 1960's and early 1970's throughout the anatomical sciences as the first minicomputers became available for laboratory work. The hope quickly arose that there would soon emerge computer-based technology for acquiring, analyzing and imaging of two- and three-dimensional information which previously could only be represented numerically. In this paper we will describe the software and hardware strategies evolved in our laboratory over the past six years to meet a broad range of quantitation and modeling applications.

BACKGROUND

One of us (DJW) (with Dr. Alan Selverston, at University of California) had the pleasure of organizing a planning session on Computer Assisted Neuroanatomy for the Division of Research Resources in 1976 in San Diego, California, USA. Seventy scientists gathered to debate the needs thought to exist for applications of computers in anatomy. It became clear that many parallel, but independent efforts were underway to study biological tissue at the gross, light or electron microscopic levels (Macagno, et. al., [4]). Data acquisition, either manually or through automated techniques, was a major concern. Alignment of serial section data required inventive new strategies. Imaging and plotting of two- and three-dimensional data arrays was recognized as a universal problem. Numerical analysis was needed for all types of data. It was also recognized that data archiving and sharing between laboratories could become commonplace. The cost of the labor involved in

preparing programs even then seemed greater than the cost of the equipment, and that distribution of software, if possible, would greatly benefit the neuroscience community. Also, it is now evident that the broader needs of brain science overlapped extensively with comparable issues in the study of objects with radiological techniques.

One product of the planning meeting was a design for a comprehensive neuroanatomical computer-based analysis system which would satisfy a wide range of needs. In our laboratory at Dallas we have directed our efforts toward creating a multipurpose system. From the outset we hoped to design the software programs and hardware with the maximum versatility for different applications in anatomy. The CARP system (Computer Aided Reconstruction Program) has been developed from this point of view. In many ways the basic host computer and programming language has not changed in the past decade. The computer memory and hardware components have become far less expensive but programming costs are greater, so that total costs probably remain similar. The major technical advance has been the introduction of the image analysis and graphics computers specifically designed for acquisition and synthesis of high-resolution video images.

The development of CARP began in earnest with the acquisition of the first commercially produced high resolution raster graphics system from Ikonas Inc. (later purchased by Adage Inc.). The unique combination of features included video digitization, a flexible color look-up table and graphics controller and a special purpose bit-slice graphics processor designed to allow a range of image analysis and sophisticated graphics operations to be done at a speed which made practical a wide range of applications in anatomy.

Three major tasks have emerged as functions of the CARP system. Morphological information is: 1) acquired by a variety of strategies, 2) manipulated by a complex data base structure, and 3) imaged by an assortment of algorithms.

HARDWARE COMPONENTS

As shown in Figure 1. the core of the system (Schlusselberg, et. al., [7] and Smith, et. al., [9]) is a host computer time shared among several input stations. We thought at the start that one-half megabyte of main memory and a 200 megabyte disk drive would be more than adequate. However, in time the programs grew increasingly complex to accommodate the variations requested by many individual users. Our current version of CARP requires a minimum of 2 megabytes of main memory with a virtual memory operating system and a minimum of 300-500 megabytes of disk space. To a great extent computers with the Unix operating system with Fortran and C languages may soon emerge as a commodity in which very complex software systems can be run by machines from many manufacturers. Our current computer host is a Data General MV/8000-II. In our experience, the AOS/VS and MV/UX environment, with its 32-bit virtual memory capabilities and hierarchical file system, has been ideal for development of large application programs.

The graphics computer is a newer concept for which few standards have emerged. The Adage 3000 raster graphics system provides for video digitization, 4 megabytes of image memory (displayable as 512 x 512 or 1024 x 1024), color coding of video output, and the microprogrammable graphics processor. Data input modules include tablet with hand-held cursor for drawing lines and points of projected images. A tablet, stepper motor stage, drawing tube, and bit map graphics terminal are used, when appropriate, to manually input data while viewing microscope slides. Finally, a high-precision film transport with movable platform, a high resolution video camera, and microfilm viewing system allows for analysis of serial electron microscopic sections.

SOFTWARE COMPONENTS

Data Input

The input data acquisition routines are divided into modules with commands suited f the operation of given hardware. TRACE is a

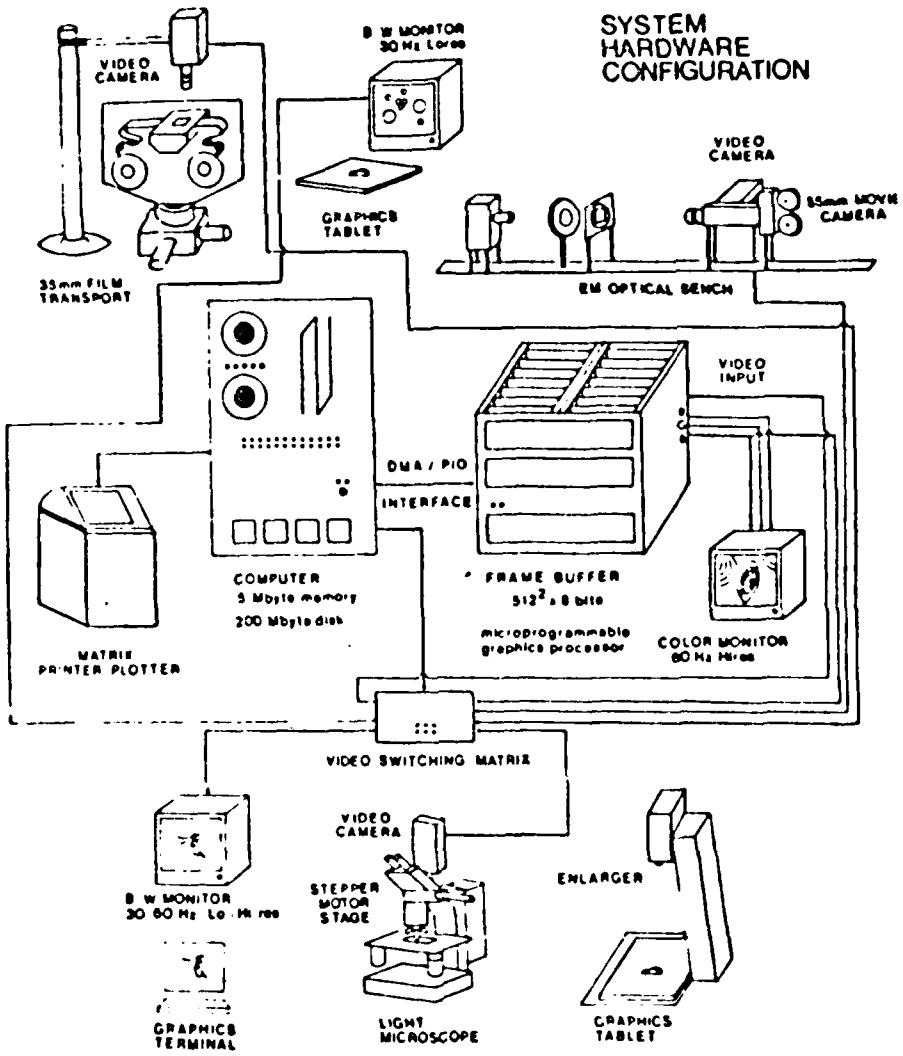


Figure 1. System Hardware Configuration.

The current computer system is built around a DATA GENERAL MV/8000-II 32-bit super-minicomputer with 6.0 megabytes of memory and two 354 megabyte disk drives, and an ADAGE 3000 Raster Graphics System with high-speed microprogrammable graphics processor in addition to a DATA GENERAL Eclipse S/130 with 0.5 megabytes of memory and one 200 megabyte disk drive. Various input stations acquire data from an enlarger, microscope, film transport and video digitizer.

set of routines which support drawing diagrams consisting of lines and points. Axes and scales are specified. Labels are applied to each line and subgroups of points. The TRACE routine allows morphometric analysis of any image represented as a histologic or photographic slide projected onto a tablet. Photographic or other images on paper can be traced. Files containing segment data can be plotted and information may be printed on line lengths, centroids and areas of closed contours. The TRACE facility has been used extensively to determine areas of cross-sections of human gross brain material or to outline brain cross-sections which will later be studied under higher power microscopy.

For radiological applications, a set of subprograms (CT for Computed Tomography, MR for Magnetic Resonance Imaging and ET for Emission Tomograpy) are ideally suited for manipulating medical image data. These routines contain provisions for reading various tape formats, displaying the image data using standard windows and acquiring perimeters around objects of interest. Manual tracing is supported in a fashion similar to TRACE; however, the tablet driven cursor is superimposed on square pixel video arrays. Identical software is used in all cases in which manually drawn lines are input.

MICRO contains routines for plotting data with a light microscope, tablet drawing tube, and bit map graphics terminal. The host computer issues command pulses to drive a stepper motor stage to specified positions. The operator inputs an origin and axis for each section of biological material. The stage can be moved randomly, directed toward positions in biological coordinates, or made to move in a grid-like pattern for systematic examination of large areas. Lines and points drawn on the tablet are converted in the host computer into floating point numbers in the biological coordinate system. Conversion routines plot lines and points on the graphics terminal. The bit mapped screen is viewed through the drawing tube. Calculations result in lines and points being superimposed on the image viewed through the eyepiece. This arrangement allows easy comparison between visual images and overlays of what has been stored in the

Systems, for a Data General Eclipse MV/8000-II 32-bit super-minicomputer with magnetic tape backup and 700 megabytes of disk storage. All data is physically stored in a binary file which is mapped into the program's virtual address space.

Data records are organized into nodes which contain information about the type of data stored and address pointers to lists of graphical data and other nodes. The structure of the graphical data list depends upon its type; for example, a segment consisting of a line points to a list of data triplets. Each triplet uses 32-bit floating point numbers to represent X, Y and Z coordinates in 3-space. Segments can include lines (e.g., tracings around the border of a sectioned object), cell positions, autoradiographic grain positions, or fiducial marks (landmarks with relatively constant positions in serial sections which are used for alignment.)

Our data file representation (Smith, et. al., [10] and [11]) of a complicated neuroanatomical object is organized in a tree-like structure which is similar to the Knuth transform of an n-ary tree (Pfaltz [5]). A tree is defined as a system of the nodes, or data packets. The tree-like structure comes from the address pointers, within the packets, which are used to determine where to find in the file the other nodes in the tree. At the top of the tree is the ROOT node which contains general information and pointers to family modes. A FAMILY node might specify, for example, how to find the location in three dimensions of all cell bodies which are part of a labeled nuclear region. Alternately, a different category of FAMILY node might point to all contours or sets of polygon strips which define a surface around a nuclear region. A variety of different node types are employed to define all the types of information encountered in neuroanatomical studies (Figure 2).

A major graphical problem exists when an object branches. In this case, it is necessary to construct a hierarchy in the database to represent the relationship among branches. Special FAMILIES of nodes are used for temporary storage and manipulation. If one branch splits into two branches, the three

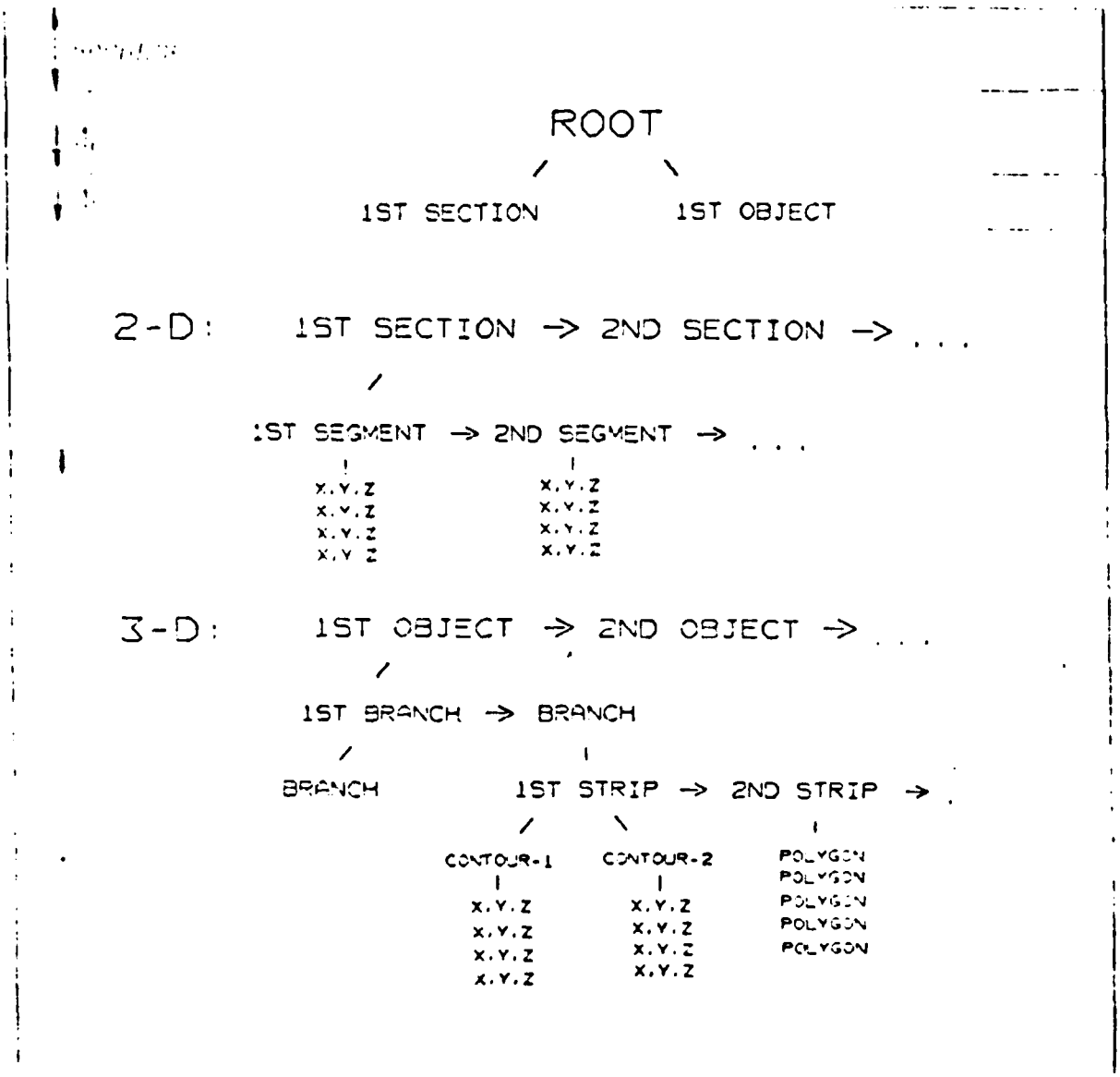


Figure 2. Hierarchy in 3-dimensional anatomic database.

A) SECTION nodes are considered siblings, and are doubly linked. SEGMENT nodes are the children of SECTION nodes, and also contain pointers to lists of three-dimensional coordinates in the data file.
B) An OBJECT node represents the complete surface description of a three-dimensional structure. OBJECT nodes are connected as siblings; their children are BRANCH nodes. BRANCH nodes connect to each other in a hierarchical fashion, depending on the branching nature of the structure. Each BRANCH node points to a family of STRIP nodes, which contain the surface description of that BRANCH. The STRIP nodes point to pairs of CONTOUR nodes and to ordered polygon lists, which are groupings of three-dimensional coordinates for vertices and normal vectors. CONTOUR nodes point to three-dimensional coordinate lists in a similar manner as SEGMENT nodes.

branches will share a contour in one of the sections. To satisfy the triangulation requirements this shared contour must be present in all three branch definitions, but should be split in two of the branches. In our data manipulation procedure, the user specifies where to split the contour because of the unpredictability of contour complexity in serially sectioned biological material.

The data management routines allow manipulation of data with use of the tree and node structure. All data is input initially from a sequence of sections and stored in portions of files located by SECTION nodes. The BUILD facility allows sections to be aligned and edited. New nodes are created to define OBJECTS which define ways of gaining access to all contours in sequence which define a continuous surface or a cluster of points. Sequential contours can be subjected to a TILE routine which creates surface definitions and determines normals of adjacent polygons to permit graphical shading routines. A QUANT facility operates on the tree-like data structure to determine volumes, surface areas, cell counts, pixel counts, graphs, bar charts and histograms.

Data Display

A goal has been to develop a full range of raster scan based video imaging capabilities. The most common two-dimensional representation of an anatomical object is a line traced around its external boundary as viewed in a serial section. These line drawings are stored as ordered vertex lists, where the first point in the list is graphically represented as a "move" (pen up), and subsequent points are "draws" (pen down). Similarly, graphical symbols are used to represent single point data such as cell positions, autoradiographic grains, synapses, etc. Symbol definitions are normalized coordinates stored in the program or data file and are translated to an anatomical coordinate position for final display. A filled region, such as a pigment-containing cell body, can be represented by center of mass, area, and intensity, when using digitized video images for data acquisition.

The graphical display of digitized information requires the definition of an anatomical window through which a part of the anatomical coordinate area is viewed. This requires clipping out portions of data that are outside of the window. The anatomical window is then mapped to a physical display device by using display coordinates which define a viewport. The window can be combined with images of objects seen in data input devices or manipulated by the user to focus in on certain regions and create summary diagrams of all data entered. The current parameters used to define this window are stored in the data file, so that it is not necessary for a user to redefine them each time a data file is examined.

There are several strategies for creating three-dimensional models of previously digitized anatomical data. By specifying certain viewing parameters and applying a perspective transformation, display coordinates are generated from vertex lists to create three-dimensional line drawings. Depth perception in these displays can be enhanced by using line intensity or thickness to represent distance from the eye or by hidden-line algorithms, which assume that each vertex list represents an opaque polygon.

Our efforts have been directed toward exploring new techniques for computer reconstruction of three-dimensional surfaces of objects which have been serially sectioned. Triangulation algorithms are used to generate an ordered polygon list from two vertex lists representing the points defining two "contours" of an object (Fuchs, et. al., [1]). An "object" in this case is usually defined as a surface bordering an anatomic region. A contour represents the external boundary of an object which has been sectioned by a plane. Contours are derived from segments in sequential sections.

There are several requirements of triangulation algorithms which must be satisfied by the database. For any given section, several segments can be digitized, representing different objects that have been sliced in that section. Each segment that belongs to one object will generate one contour. Techniques have been

developed to specify which contours belong to the same object in different sections. If an object branches, it will generate more than one contour in each section. Special triangulation algorithms are needed to generate a proper set of polygons for branching surfaces when given more than two contours as input. It is also necessary to keep track of the branching hierarchy in biological objects, since this may contain important structural and functional information. The simplest way to accomplish this is to store a set of sequential contours as one branch, which represents one portion of a three-dimensional surface. Each branch serves as input to a triangulation algorithm, since it only contains one contour per section. By manipulating portions of surfaces in this way it is possible to represent surfaces of considerable complexity by raster scan computer graphics.

For data display, there exists a package of graphical output routines that operates on the data structure to output graphical primitives. In 2D mode, vector lines, points and symbols are plotted on a variety of devices. In 3D mode, lines can be assigned color and intensity and line width to achieve depth shading. As plotting progresses in the frame buffer the Z depth positions can be stored and compared pixel by pixel. This technique allows calculation of hidden surfaces and transparencies. Illumination can be calculated from information about normals to the polygons which define the surface. The Gouraud [2] method of producing shading and illumination optimizes use of pseudocolor and high speed of microcode programs. The Phong [6] method uses a full twenty-four bits for full color representation. A more satisfactory view of a solid plastic-like surface is produced but at the expense of cpu time. To produce a high quality image, a user needs to position calculated light sources, determine reflectance values and set many parameters, including eye position and distance, much as with a conventional photographic or illustration process.

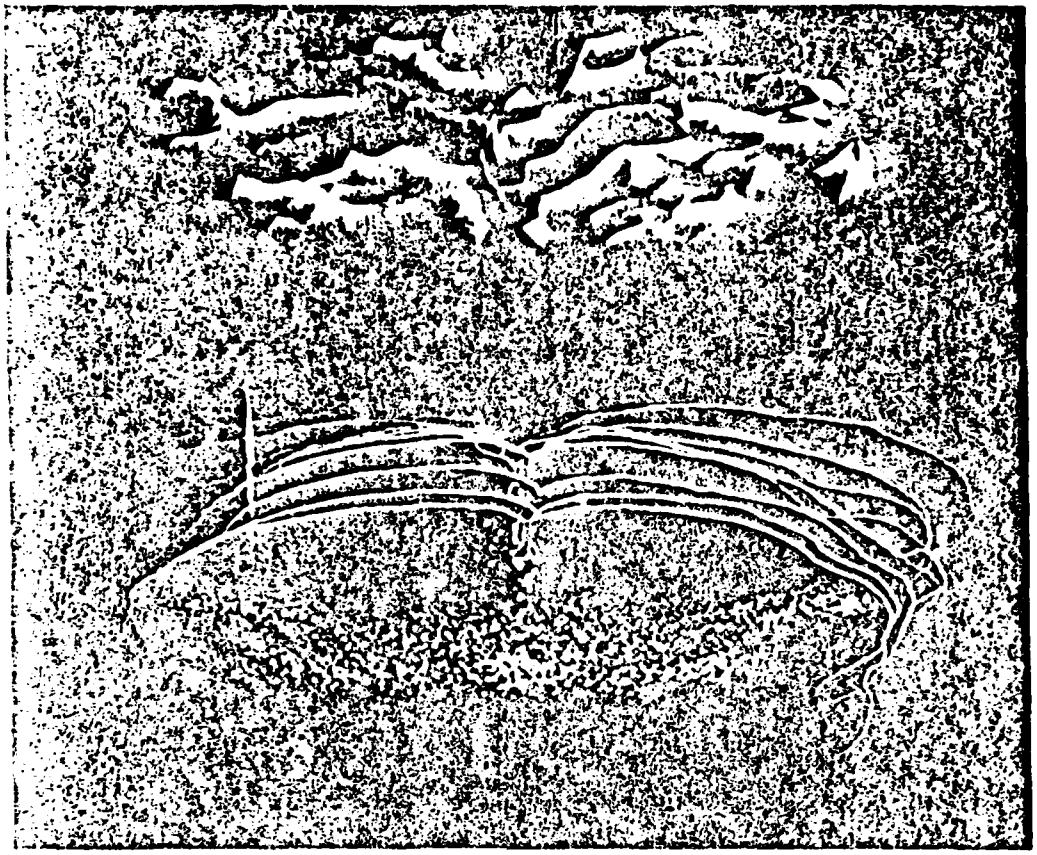
These solid-body modeling techniques have been used in a number of projects involving traditional anatomical questions. Studies of human disease processes such as Schizophrenia and Parkinson's Syndrome have implicated a part of the brain known as the

substantia nigra. Cells in this region appear to manufacture too much dopamine (Schizophrenia) or too little (Parkinsons's); this may be related to increased or diminished production per cell, although it appears more likely that it is due to an actual change in the number of cells. It has long been recognized that there is a steady decrease in the number of cells in the substantia nigra with aging. A major undertaking has been to use the computer to count these cells in a wide range of human patients and other species such as rats and mice. It may well be that regional changes in these dopamine cell populations are responsible for the presentations of these diseases. The CARP system has been expanded to include a variety of special display and quantitation capabilities for arbitrary cell populations. (Figure 3).

Another very complex problem in general biological reconstructions involved an anatomic analysis of a renal nephron from the kidney of the little skate, Raja erinacea. This elasmobranch fish is able to maintain a serum osmolality nearly that of seawater by concentrating urea from its glomerular filtrate. A highly convoluted grouping of parallel tubules wrapped in a cellular sheath may act as a physiological countercurrent system. CARP was used to reconstruct this tubular system from a set of data that included over 120 sections and over 3000 perimeters (Lacy, et. al., [12]). The peritubular sheath acts to separate adjacent nephronal systems and allow localized solute gradients (Figure 4). The discovery of a unique countercurrent flow system within each nephron resulted directly from the reconstructed visualization of the tubules.

In a more recent project, measurements of regional brain volume were acquired from Computed Tomography data and used to compare levels of brain atrophy in aging patients. A dividing plane passing though the pineal gland and the floor of the fourth ventricle and perpendicular to the midline plane was constructed to create anterior and posterior compartments in the cranial cavity. Perimeters around the brain and inner lining of the skull were separated into left and right, anterior and posterior regions (Figure 5). Volumes of these regions were computed both

3.1.12



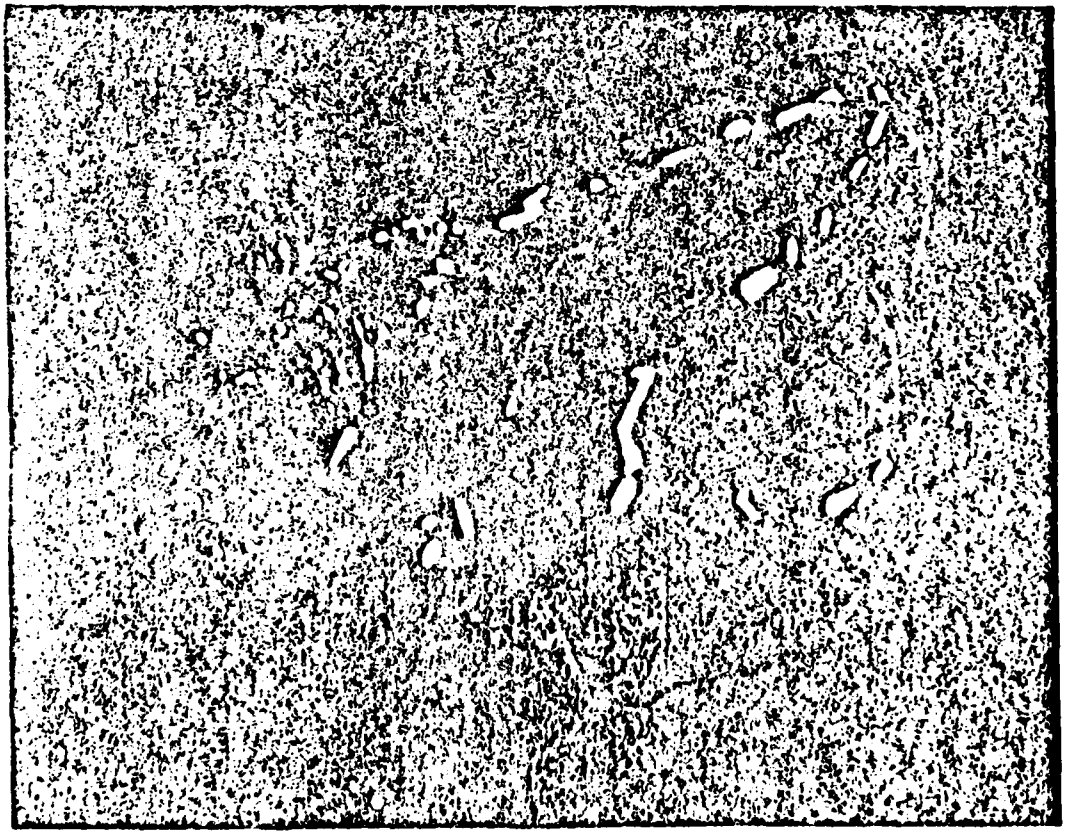


Figure 3. Dopamine cells from substantia nigra of mouse brain.

At the bottom is a rostral view of a series of sections through a mouse midbrain and pons. Yellow spheres mark the locations of cells containing dopamine pigment in the substantia nigra which were entered through the light microscope. Above the sections is a cell density distribution plot with the height of the color-coded surface representing the number of cells directly beneath.

Figure 4. Reconstruction of portion of nephron from the kidney of the elasmobranch little skate, Raja erinacea.

Data consists of approximately 100,000 polygons generated from 3100 tubule profiles through 125 serial slices. This image shows a cellular sheath surrounding the specialized bundle of five tubules arranged in a parallel array. This bundle is thought to be responsible for the ability of the skate kidney to regulate body fluids in sea water.

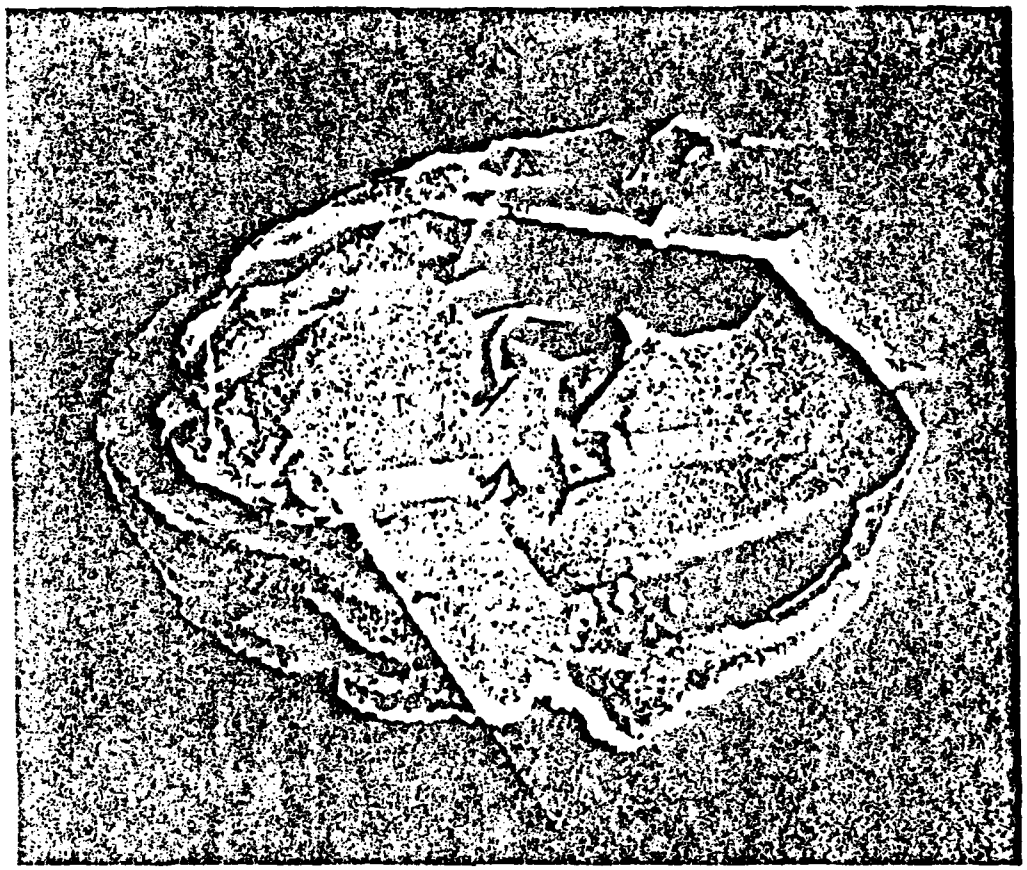
Regional brain indices were calculated by the ratio of brain volume and cranial volume for each region. The regional brain indices were considered to be the normalized index of atrophy for comparison between patients.

Command Processing

A more recent realization has been the need to organize in a sophisticated way the inumerable elementary software operations to execute database input and output. Our strategy has been to organize this large highly interactive program by designing software modules which perform distinctive tasks and allowing the user to link them together using a straightforward command-driven interface or to build command procedures which use a LISP- and C-like syntax to perform a more complex task.

These procedures are written in a CARP command language which supports a variety of standard language constructs - looping, conditional tests, arithmetic operations, scientific mathematical functions and string manipulations. Internal symbol table management routines allow definition of CARP numeric and string variables. These features contribute to the ability of a general user to identify the repetitive keystroke and interactive operations and reduce them in a tailored command procedure which may provide either a menu-driven or command-driven interface

3.1.14



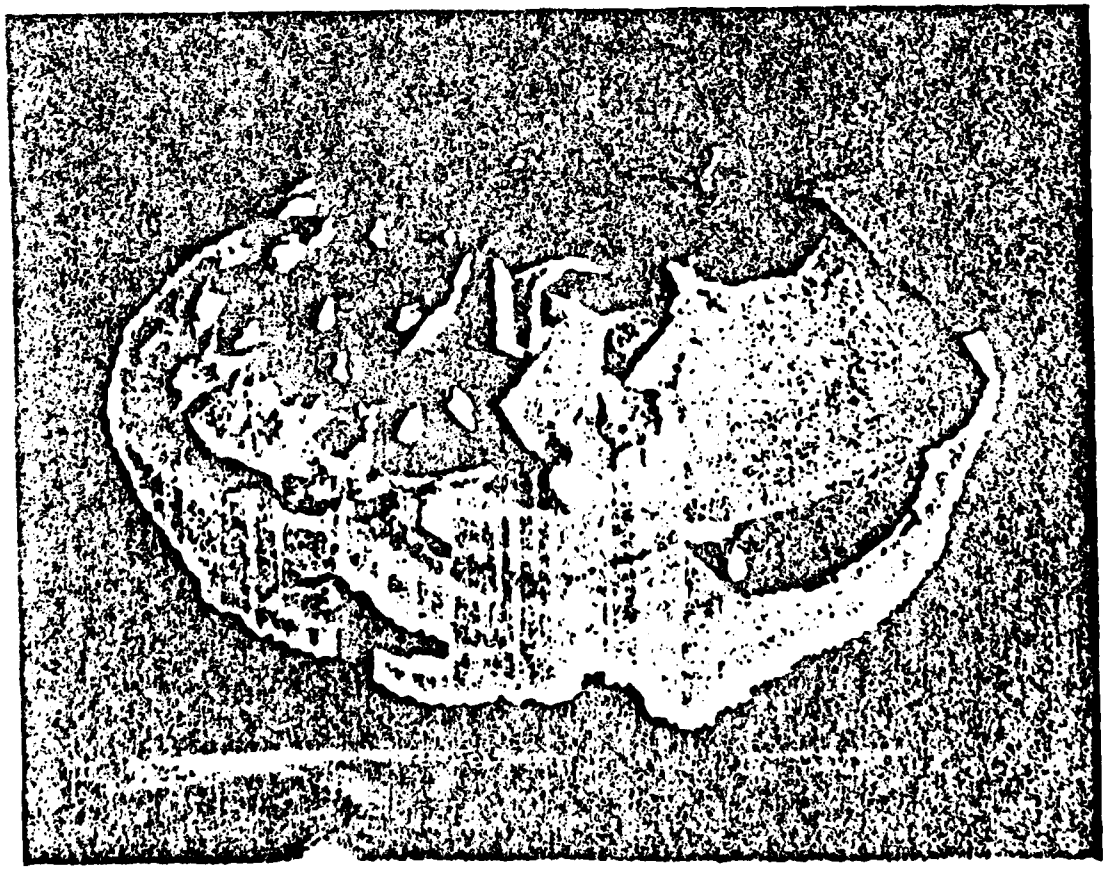


Figure 5. Regional brain volume determination in human aging patients.

A) A dividing plane passes through the pineal gland and the floor of the fourth ventricle perpendicular to the midline plane which separates the brain and cranial cavity into anterior and posterior compartments.
B) Regions are color-coded to show left and right, anterior and posterior divisions. Base of skull is modelled using cuberille surface generation techniques.

DISCUSSION

Our major goal has been to design a flexible general-purpose system for quantitative analysis and generation of high resolution three-dimensional images of biological material that has been serially sectioned either mechanically as in traditional anatomical studies, or electronically as in contemporary radiological imaging techniques. It is clear that there is a great deal of commonality in the software requirements of the broad range of applications we have encountered. Data can be input from a variety of devices, and because we use absolute (biological) coordinates stored as floating point numbers, data values can have a wide range (sub-micron to macroscopic). The absolute coordinate system also allows data from the same material to be analyzed with different modalities such as magnified projections onto a tablet or views from a light microscope.

Because of the complex and unlimited nature of biological objects, the database design had to be tailored to provide efficient storage and access. The database routines also had to be adaptable to a wide variety of experimental paradigms such as surface reconstruction, cell and grain counting, morphometric analysis, and regional density. Database design has constituted a major programming effort, at times consuming 90% of our effort, and has proven invaluable to the flexibility of the system.

Another valuable asset has been the implementation of various imaging strategies which can utilize separate components within the hierarchical database to allow visualization of relationships

within complex biological objects. The need has been to make routines be highly interactive, allowing the user to display the specific portions of three-dimensional objects of interest. The same routines are used to condense large amounts of quantitative information from multiple three-dimensional objects into color-coded 3D graphical images, which correspond to distributions of densities within 3D objects. This allows direct comparison of asymmetries within a single object, paired objects, or objects from different subjects.

Finally, as we evolved through multiple revisions of software, the need became apparent for an optimal organization of primitive operations, basic commands and configured command procedures in a variety of responsive menus and user interfaces.

ACKNOWLEDGEMENTS

We acknowledge the support from grant numbers AA-390, NIAAA 1F32-5198, DA-2338 and the Biological Humanics Foundation.

REFERENCES

1. Fuchs, H., Kedem, Z.M. and Uselton, S.P. (1977). Optimal Surface Reconstruction from Planar Contours. Commun ACM., 20, 693-702.

2. Gouraud, H. (1971). Continuous Shading of Curved Surfaces. IEEE Trans Comput., 20, 623-629.

3. Kernighan, B.W. and Ritchie, D.M. (1978). The C Programming Language. Prentice-Hall, New Jersey.

4. Macagno, E.R., Levinthal, C. and Sobel, I. (1979). Three-dimensional Computer Reconstruction of Neuronal Assemblies. Ann Rev Biophys Bioengng., 8, 323-351.

5. Pfaltz, J.L. (1977). Computer Data Structures. McGraw Hill, New York.

6. Phong, B.T. (1975). Illumination for Computer Generated Pictures. Commun ACM., 18, 311-317.

7. Schlusselberg, D.S., Smith, W.K., Lewis, M.H., Culter, B.G. and Woodward, D.J. (1982). A General System for Computer-Based Acquisition, Analysis and Display of Medical Image Data. ACM 1982 Conf Proc., ACM Annual Meeting.

8. Schlusselberg, D.S., Smith, W.K., Culter, B.G. and Woodward, D.J. (1982). A Computer System for Semi-Automatic Cell Recognition in Neuroanatomic Studies. Soc. Neurosci. 1982 Abstract 182.9.

9. Smith, W.K., Schlusselberg, D.S., Woodward, D.J. (1981). A Computer System for Neuroanatomical Data Acquisition, Analysis and Display. Soc. Neurosci. 1981 Abstract 135.18.

10. Smith, W.K., Schlusselberg, D.S., Culter, B.G., Woodward, D.J. and Lacy, E.R. (1983). Hierarchical Database Design for Biological Modeling. NCGA 1983 Conf Proc., 106-116.

11. Smith, W.K., Schlusselberg, D.S., Culter, B.G. and Woodward, D.J. (1983). A Database Structure for Three-Dimensional Reconstruction of Neuroanatomical Objects. Soc. Neurosci. 1983 Abstract 106.5.

12. Lacy, E.R., Reale, E., Schlusselberg, D.S., Smith, W.K. and Woodward, D.J. (1985). A Renal Countercurrent System in Marine Elasmobranch Fish: A Computer-Assisted Reconstruction. Science, 227, 1351-1354.

THE JOURNAL OF COMPARATIVE NEUROLOGY 233:350-364 (1985)

Rat Medulla Oblongata. IV. Topographical Distribution of Catecholaminergic Neurons With Quantitative Three-Dimensional Computer Reconstruction

MADHU KALIA, DONALD J. WOODWARD, WADE K. SMITH AND KJELL FUXE
Department of Pharmacology, Jefferson Medical College of Thomas Jefferson University, Philadelphia, PA 19107 (M.K.), Department of Cell Biology, University of Texas Health Science Center at Dallas, Dallas, Texas 75235 (D.J.W., W.K.S.), and Department of Histology, Karolinska Institutet, Stockholm, Sweden (K.F.)

ABSTRACT

We examined serial 40 μm vibratome, immunoperoxidase-stained sections of the medulla with tyrosine hydroxylase (TH), dopamine-beta-hydroxylase (DBH), and phenylethanolamine N-methyltransferase (PNMT) antisera followed by Nissl staining to locate catecholaminergic neurons in cytoarchitectonic regions followed by a three-dimensional (3D) computer reconstruction of these cell groups to determine their spatial organization. Overlay drawings of low and high power photomicrographs showing cell bodies and nuclear boundaries were entered into a digital computer storage system. Every section in the series was plotted to yield an accurate representation of regional densities of cells and location of nuclei, as revealed by two-dimensional plots of individual sections as well as three-dimensional plots of groups of sections. Data files were scanned in a number of ways to obtain total cell counts of TH-, DBH-, and PNMT-immunoreactive cells within a designated area or cell counts of only one type of immunoreactive cell. This combination of data manipulation produced the following results: (1) *A1 group* is a homogeneous population of noradrenergic neurons at levels caudal to the obex, and at the obex it is mixed with adrenergic cells. The dimensions of the A1 cell group are 1.3 × 2.7 mm, extending from −2.5 to +0.2. Part of this cell group lies in the lateral reticular nucleus. (2) *A2 group* is not purely noradrenergic as previously suspected. It is a very mixed cell group containing mainly dopaminergic neurons in the area postrema (periventricular region) and the dorsal motor nucleus of the vagus, mainly noradrenergic neurons in the medial subnucleus of the nucleus of the tractus solitarius (nTS), mainly adrenergic neurons in the dorsal strip and dorsal subnucleus of the nucleus of the tractus solitarius, and a mixture of all three catecholaminergic neurons in the other subnuclei of the nTS. The dimensions of this group are 0.4 × 3 mm extending from −2.7 to +0.3. (3) *C1 group* is a homogeneous population of adrenaline cells extending from +1 to +2.5 with dimensions of 1.5 × 1.5 mm and consisting of scattered neurons some of which occupy the gigantocellular reticular nucleus. (4) *C2 group* is a homogeneous population of adrenaline neurons extending from +1 to +3 with dimensions of 2.5 × 3 mm. Accurate visual imaging and quantitation of the spatial organization of medullary catecholaminergic neurons within the classical anatomical framework of cytoarchitecture provides an enhanced comprehension of the organization of this region of the central nervous system.

Key words: dopamine, noradrenaline, adrenaline, tyrosine hydroxylase, dopamine-β-hydroxylase, phenylethanolamine N-methyltransferase

Accepted October 4, 1984

Address reprint requests to Madhu Kalia, M.D., Ph.D., Professor of Pharmacology, Room 329, Jefferson Alumni Hall, Thomas Jefferson University, 1020 Locust Street, Philadelphia, PA 19107.

In the past, localization studies of catecholamine-containing neurons in the central nervous system have relied on the use of formaldehyde histofluorescence and immunofluorescence techniques. A number of investigations of catecholaminergic neurons have been conducted on the medulla oblongata since this is an important region for the integration of central nervous system mechanisms involved in autonomic, cardiovascular, and respiratory functions. The monoaminergic neurons in the brain described by Dahlström and Fuxe ('64) were designated by the letter "A" and classified as noradrenaline-containing neurons. These studies were based on the Falck-Hillarp technique ('62). A second population of monoaminergic neurons, the "C" cell groups (Hökfelt et al., '73), were identified as adrenaline-containing neurons. Details of the "A" and "C" catecholaminergic cell groups and their relationship to cytoarchitectonic boundaries in the medulla oblongata have been presented in the two companion articles preceding this paper (Kalia et al., '85a,b).

The question of where these catecholaminergic neurons are located in the brain stem has been addressed by a number of previous investigators, but the use of fluorescence microscopy has not permitted a simultaneous analysis of cytoarchitecture and immunoreactivity. The immunoperoxidase method of Sternberger ('79) permits light microscopic visualization of immunoreactive nerve cell bodies, nerve fibers, and preterminal processes and allows subsequent staining of this material with Nissl stain to precisely determine cytoarchitectonic boundaries of the region so that the correlation between cytoarchitectonically distinct cell groups and the catecholaminergic neurons can be clearly defined. In order to correlate cytoarchitecture with the location of catecholaminergic neurons in this region of the brain stem we have employed the immunoperoxidase method of Sternberger ('79) along with a method of overlay drawings showing the location of immunoreactive neurons within precise subnuclear groups of the dorsal medulla (Kalia et al., '84). In this study we have combined four techniques: (1) bright-field immunoperoxidase technique on serial vibratome sections; (2) overlay drawings of photomontages of the medulla oblongata from the different catecholaminergic cell groups (A1, A2, C1, and C2); (3) Nissl counter-staining of the same sections; and (4) three-dimensional computer reconstruction (Smith et al., '82).

Identification of catecholaminergic neurons by means of immunocytochemistry requires as a first step localization of the catecholamine-synthesizing enzymes. The presence of tyrosine hydroxylase (TH) indicates the existence of dopaminergic, noradrenergic, and adrenergic neurons; the presence of dopamine-beta-hydroxylase (DBH) indicates the presence of noradrenergic and adrenergic (but not dopaminergic) neurons; the presence of phenyl ethanolamine N-methyl transferase (PNMT) indicates the presence of adrenergic (but not dopaminergic and noradrenergic) neurons. Therefore if a group of neurons shows immunoreactivity to all three catecholamine-synthesizing enzymes (TH, DBH, and PNMT), then those neurons can be considered to be adrenaline-containing; if a group of neurons shows immunoreactivity only to TH and DBH, then those neurons can be considered to be noradrenaline-containing; and a group of neurons showing only TH (and not DBH and PNMT) immunoreactivity is most probably dopaminergic.

It is necessary to have information about all three catecholamine-synthesizing enzymes in serial sections to be able to determine the neurochemical identity of catecholaminergic neurons. In addition, we need to visualize the location of all three catecholamine-synthesizing enzymes simultaneously so that we can see which cell groups contain which enzyme. The two preceding companion papers (Kalia et al., '85a,b) contain detailed descriptions of the location of catecholamine-synthesizing enzymes in various regions of the medulla oblongata of the rat. This paper uses that data on TH-, DBH-, and PNMT-containing neurons and provides us with visual images so that the spatial organization of neurons containing catecholamine-synthesing enzymes can be accurately determined. The three-dimensional computer reconstruction method for producing these visual images proved to be most useful in this essential step in our analysis which required the simultaneous visualization of all three groups of immunoreactive neurons.

The purpose of this study was to quantitatively analyze the rostrocaudal and mediolateral location of catecholaminergic cell groups in the medulla oblongata. In addition, it was of interest to determine the overlap and/or continuity between the dorsal (A1 and C1) and the ventral (A2 and C2) cell groups, two of which contain adrenaline (C1 and C2), the two others containing noradrenaline (A1 and A2). The spatial organization of these different groups of catecholaminergic neurons within the cytoarchitectonic boundaries of the brain stem was investigated in this study to uncover the functional significance of these observed morphological and immunocytochemical distinctions.

METHODS

Histological sections through the brain stem obtained from experiments described in the preceding two articles (Kalia et al., 85a,b) were evaluated as follows: The sections containing monoamine immunoreactive neurons were photographed at 2× magnification with an Olympus Vanox Photomicroscope using bright-field illumination with an open field aperture and condenser setting to enhance the contrast. The total image of the section was photographed within one frame, and high contrast prints were made. (Fig. 1A). From the photographic prints, overlay drawings of the sections were carefully made, the location of neuronal profiles was marked by triangles, and the major anatomical landmarks (nuclear boundaries and fiber tracts) were drawn as straight lines (Fig. 1B). The number and location of cells was monitored by continuously checking the section under the microscope during the drawing procedure.

Abbreviations

ap	area postrema
DBH	dopamine beta hydroxylase
dmnX	dorsal motor nucleus of the vagus
dnTS	dorsal nucleus of the nucleus of the tractus solitarius
dPSR	dorsal parasolitarius region
ds	dorsal strip
LRt	lateral reticular nucleus
ml	medial lemniscus
mlf	medial longitudinal fasciculus
mnTS	medial nucleus of the nucleus of the tractus solitarius
ncom	commissural nucleus of the nucleus of the tractus solitarius
nTS	nucleus of the tractus solitarius
PGi	paragigantocellular reticular nucleus
PNMT	phenylethanolamine N-methyl transferase
PVR	periventricular region
TH	tyrosine hydroxylase
TS	tractus solitarius

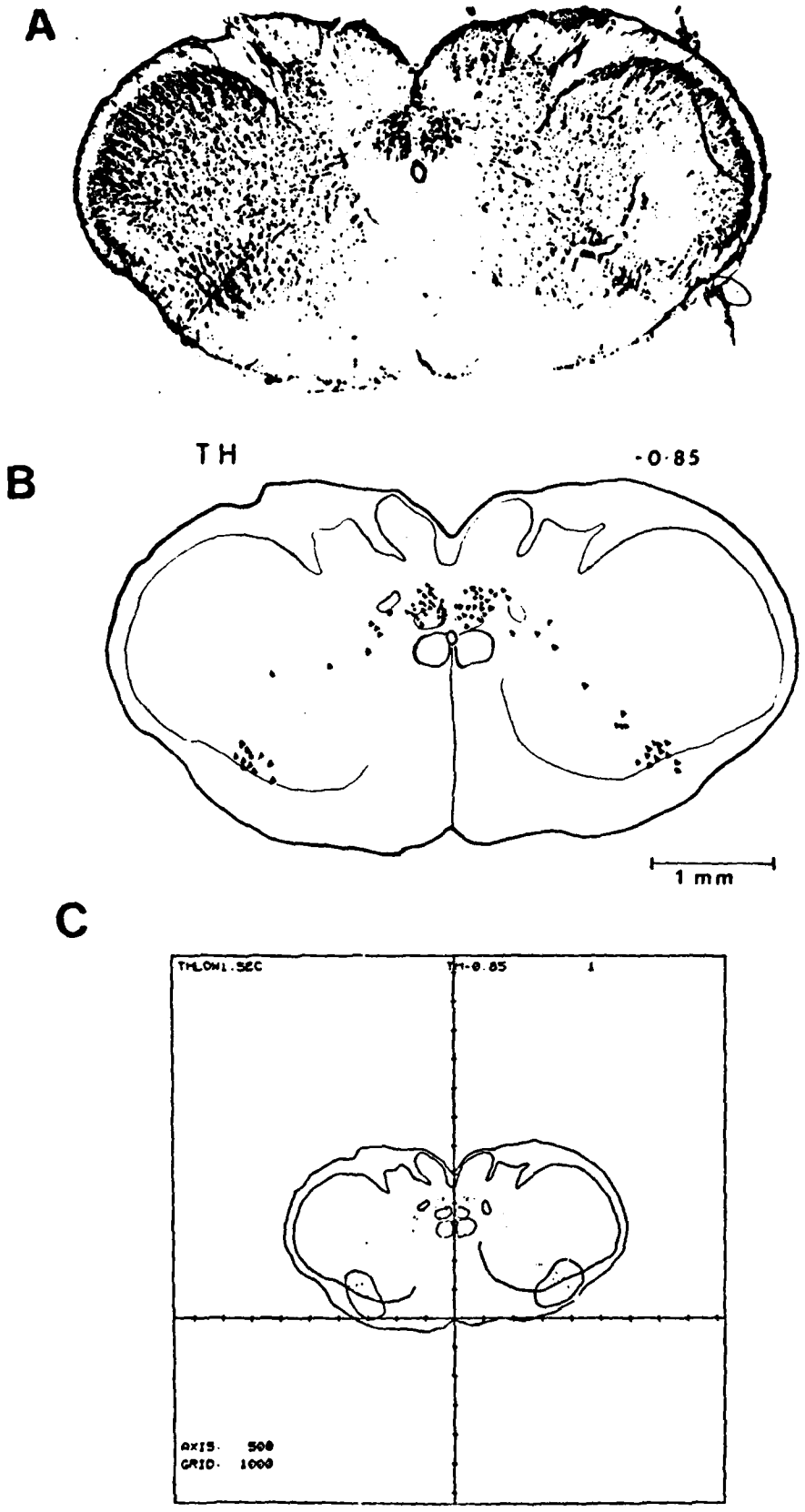


Fig. 1. A. Low magnification photomicrograph of the caudal medulla oblongata at a level −0.85 mm caudal to the obex, showing TH-immunoreactive cells. B. Overlay drawing of a section in A. Filled triangles mark the location of TH-positive cells. Major landmarks are indicated as solid lines. C. Computer reconstruction of the same section. The lateral reticular nucleus is indicated in the ventral region. Grid marks on the X and Y axes = 500 μm.

Computer reconstruction

The overlay drawing was made to fit the surface of a digitizing tablet. The drawing was laid on the digitizing tablet, and an origin and axis were determined as follows. In sections caudal to the obex, the lowest point of the central canal was taken as the origin (a) and the highest point of the area postrema on the surface of the medulla vertically above it in the region of the area postrema was taken as the second coordinate point (b) to define an axis on the midline. The magnification of the section was also recorded. The provision of two coordinate points enabled definition of an axis on the midline through which alignment adjustments could be made and through which the sections could be rotated in various planes during the computer data manipulation and processing. This procedure allowed the information to be input rapidly at low power magnification. An outline of the total section as well as lines delimiting the borders of major regions, particularly fiber tracts and nuclei, were input as separate line segments for each section in the series (Fig. 1C). In addition, points corresponding to positions of cell bodies were input into the computer through a digitizing tablet and stepping motor stage. Individual labels were assigned to groups of cells within individual subnuclei so that a subsequent analysis of a single cell group could be performed. Every section stained with the same antibody was drawn sequentially, and the rostrocaudal position of the section was marked. Graphic system software allowed the two-dimensional plots of individual sections as well as groups of sections to be visualized together in one image. The data files could be manipulated and scanned in a variety of directions to create an enhanced appreciation of the location of cell bodies in relation to one another (Smith et al., '81; German et al., '82,'83; Reis et al., '82; Schlusselberg et al., '82a,b).

These procedures provided us with newly synthesized visual summaries of all three catecholamine-synthesizing enzyme-containing cell groups (TH, DBH, and PNMT) and enabled us to accurately determine the presence or absence of any of these enzymes in any region of the medulla. This method of simultaneous visual imaging provided us with accurate information about the location of catecholaminergic cell populations in the brain stem. This was particularly important in the A2 cell group where the population contains all three types of catecholaminergic neurons, and therefore the analysis required the accuracy provided by the computer.

RESULTS

The main purpose in this paper was to convert information about the location of catecholamine-synthesizing enzymes into data regarding the position of different catecholamines in the medulla oblongata. This required analysis of catecholamine-synthesizing enzyme-containing neurons in the medulla by means of the computer with the aim of visualizing the spatial location and distribution of different groups of dopamine-, noradrenaline-, and adrenaline-containing cells.

The distribution of catecholamine-synthesizing enzyme-containing cells was plotted in the rostrocaudal and mediolateral direction at low magnifications and in relation to cytoarchitectonically distinct subnuclei in the dorsal medulla at high magnifications.

The following levels were examined at low magnification: PNMT −0.8, −0.7, −0.5, 0.1, 1.6, 1.85, 2.05; DBH −0.75, −0.3, −0.25, 0.05, 1.31, 1.5, 1.9; TH −0.85, −0.6, −0.4, −0.15, 0.05, 0.3, 1.3, 1.75, 2.2, 2.8. At high magnification of the dorsal medulla (including the nucleus of the tractus solitarius) the following rostrocaudal levels were examined: (a) levels ranging from 0.1 to 0.2 (TH 0.1, DBH 0.15, PNMT 0.2); and (b) levels ranging from −0.25, −0.15 (TH −0.25, DBH −0.2 and PNMT −0.15).

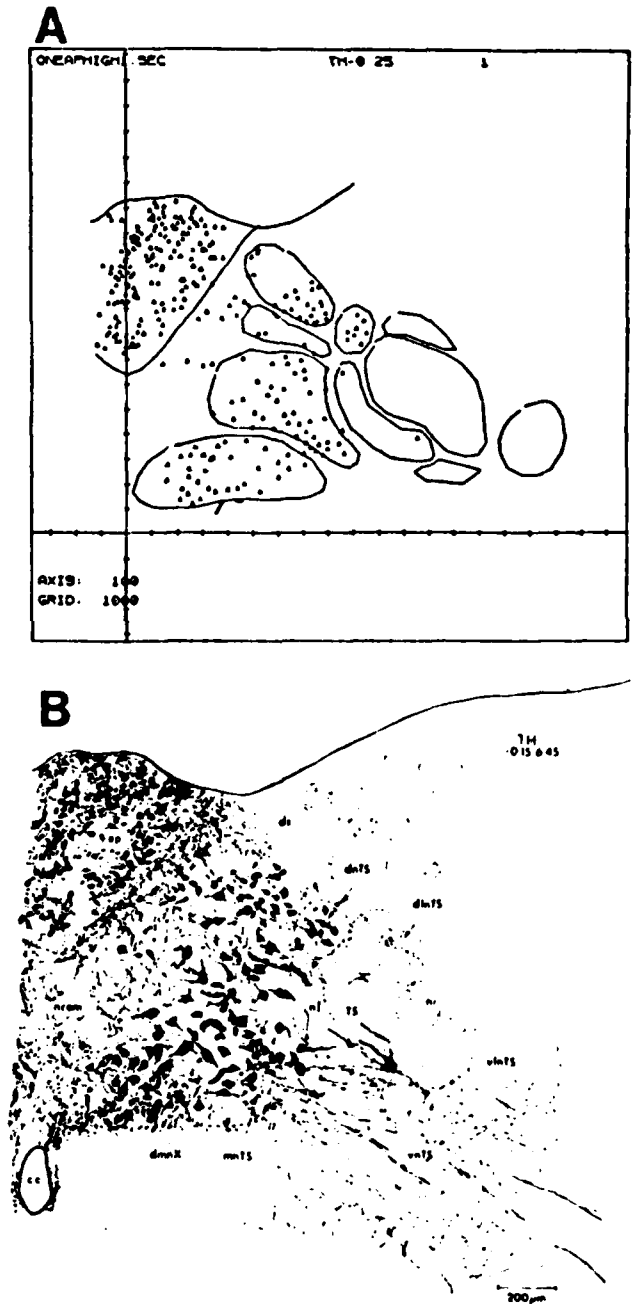


Fig. 2. A. Line drawing of TH immunoreactive neurons, nerve fibers, and preterminal processes in the dorsal medulla at a level −0.15 mm caudal to the obex. This drawing was made from an overlay of a photomontage. B. Computer plot of the same section. Cell bodies are indicated as filled triangles and major nuclear boundaries are plotted. Grid marks on the X and Y axes = 100 μm.

neurons showing TH immunoreactivity are coded green, DBH immunreactivity is coded blue, and PNMT immunoreactivity is coded white in this analysis.

Visualization of all the catecholaminergic cell groups in the medulla

All three types of catecholaminergic neurons (dopaminergic, noradrenergic, and adrenergic) show TH-positive immunoreactivity. Therefore we examined serial sections at low magnification showing only TH-positive cells (Fig. 2). Figure 3A and B show reconstructions of sections in the coronal plane showing TH-positive cells. Figure 3A is a view from the rostral side and Figure 3B is a view from the caudal side. In both figures the sections are viewed at a tilt of 30°. Figure 3A shows the distribution of the different catecholamine cell groups; in the foreground the C1 and C2 cell groups are located in the ventral and dorsal regions, respectively. In the background the A1 and A2 cell groups are visible in the central and dorsal regions, respectively. Figure 3B, which is a view from the caudal aspect, shows the A1 (ventral) and A2 (dorsal) cell groups in the foreground. The complexity of the A2 cell population is immediately obvious.

Figure 4A and B show sagittal views of TH-positive cell groups. These lateral views were produced with a 30° tilt. Figure 4A shows the caudal sections in the foreground and 4B shows the rostral sections in the foreground. Four major observations can be made from this figure: (1) the dorsal A2 and C2 cell groups do not appear to be a continuous column of cells; (2) the dorsal A2 cell group shows a distribution pattern that is very different from the dorsal C2 cell group; (3) the ventrally located A1 and C1 cell groups form a continuous column of cells; and (4) the rostrally located C1 cell group has a distribution pattern that contains scattered cells and is thus very different from the A1 cell group which contains compactly arranged cells.

Figure 5 shows reconstructions of coronal sections showing the location of all three catecholamine-synthesizing enzymes (TH, DBH, and PNMT). Figure 5A shows a three-dimensional reconstructin of a view from the caudal side (without any tilt). The most obvious result that this imaging gives is the location of two bilaterally symmetrical cell groups on the ventral medulla, two bilaterally symmetrical cell groups on either side of the midline in the dorsal medulla, and a third group of cells in the midline (in the region of the area postrema). Figure 5B shows the same group of sections with a 30° tilt. Now immediately, five populations of cells come into view: (1) a caudally located ventral bilateral cell group (A1); (2) a rostrally located ventral bilateral cell group (C1); (3) a caudally located dorsal bilateral cell group (A2); (4) a rostrally located dorsal bilateral cell group (C2); and (5) a caudally located unpaired cell group in the midline. This figure shows the usefulness of this method of visual imaging in revealing the precise location of these different populations of cells.

Figure 6 is a three-dimensional reconstruction of a series of high magnification drawings of the dorsal medulla showing the location of all three catecholamine-synthesizing enzymes in the various subnuclei of the nucleus of the tractus solitarius and adjacent regions of the dorsal medulla.

Analysis of dopamine-containing neurons in the A2 cell group. These are cells showing TH-positive (coded green) immunoreactivity, and no DBH (coded blue) or PNMT (coded white) immunoreactivity. Figures 6 and 7 enable us to visualize three regions of the dorsal medulla containing dopamine (D) neurons: (1) the medial and dorsal part of the area postrema (ap), (2) the caudomedial pole of the dorsal motor nucleus of the vagus (dmnX), and (3) the periventricular region (PVR). These neurons are located at levels caudal to the obex (Figs. 3A,6). Notice the mixture of TH-, DBH-, and PNMT-containing neurons in the other subnuclei of the nTS.

Analysis of noradrenaline neurons in the A2 cell group. These are cells showing TH (green) and DBH (blue) but no PNMT (white) immunoreactivity. The analysis of these sections at a glance shows the following. Area postrema contains noradrenaline neurons in the dorsal and medial part (note that this population appears green (TH) and blue (DBH). The second population of cells in the area postrema consists of adrenaline-containing cells which are labelled with TH, DBH, and PNMT, shown here as green, blue, and white cells. The dorsal motor nucleus of the vagus is heavily filled with dopamine-containing cells shown here in green (Figs. 6,7). Note the absence of blue and white cells in this figure. The medial nTS appears to contain primarily noradrenaline neurons. The neurons appear green and blue and not white. The dorsal parasolitarius region (dPSR), which consists of the region just dorsal to the medial nTS, contains primarily adrenaline-containing cells (green, blue, and white). This also appears to be a mixed group with adrenaline- and noradrenaline-containing cells. The dorsal strip region (ds), which contains green, blue, and white cells, appears to be predominantly an adrenaline-containing cell group since the most dense population of cells is stained white. The dorsal nucleus of the tractus solitarius (dnTS) also appears to contain primarily adrenaline-containing neurons. The intermediate nucleus (nI) located between the TS and the rest of the nTS complex also appears to contain adrenaline neurons although there is definitely some degree of mixing. The TS itself contains an occasional adrenaline neuron.

Analysis of adrenaline-containing neurons in the A2 cell group. These are neurons staining with all three catecholamine synthesizing enzymes: TH, DBH, and PNMT. However, since neither D nor NA neurons stain with PNMT, the presence of PNMT can be considered to be specific for A neurons (Goldstein, '72). These neurons are coded white in this series. Figure 6 shows the adrenaline cell group. Figure 6 shows a caudal and rostral view of the adrenaline neurons. A new population of adrenaline neurons in the ds and dnTS which is quite separate from the previously defined (Dahlström and Fuxe, '64) A2 group can be seen in the foreground.

In summary, at the level of the obex in the A2 cell group there is a mixture of all three CA neurons. The adrenaline neurons are located in the ds and dnTS, the noradrenaline neurons are located in the ap, mnTS, and ncom, and the dopaminergic neurons are located in the ap, caudomedial dmnX, and PVR. The A2 cell group extends from −2.7 to +0.3 mm and has dimensions of 0.4 × 3 mm.

Coronal sections in which the distribution of DBH-immunoreactive neurons is plotted are shown in Figure 8. This catecholamine-synthesizing enzyme (DBH) is present in noradrenaline- and adrenaline-containing neurons. Thus, Figure 8 shows the location of the combined population of adrenaline and noradrenaline neurons. Figure 8A is a view from the caudal side at a 30° tilt and shows the presence of two paired cell groups in the foreground (caudal): the A1 cell group in the ventral region and the A2 cell group in the dorsal region. Notice the absence of DBH-positive cells

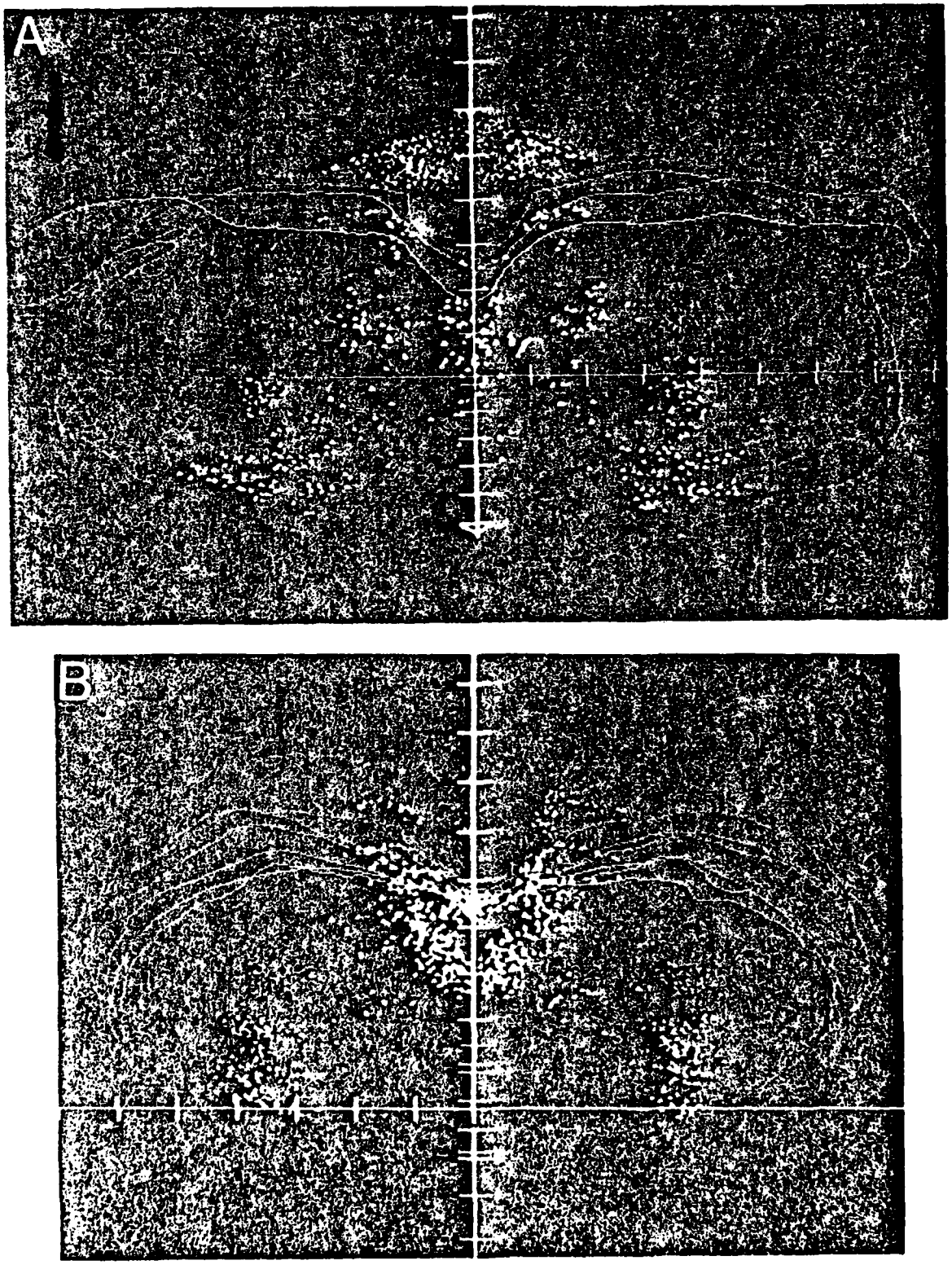


Fig 3 A Three-dimensional reconstruction of TH immunoreactive neurons in the medulla oblongata. TH positive cells are indicated as green triangles View is from the rostral side, with the caudal end tilted 30° upward B Same sections as shown in A, viewed form the caudal side, with the rostral end titled 30° upward Grid marks = 500 μm

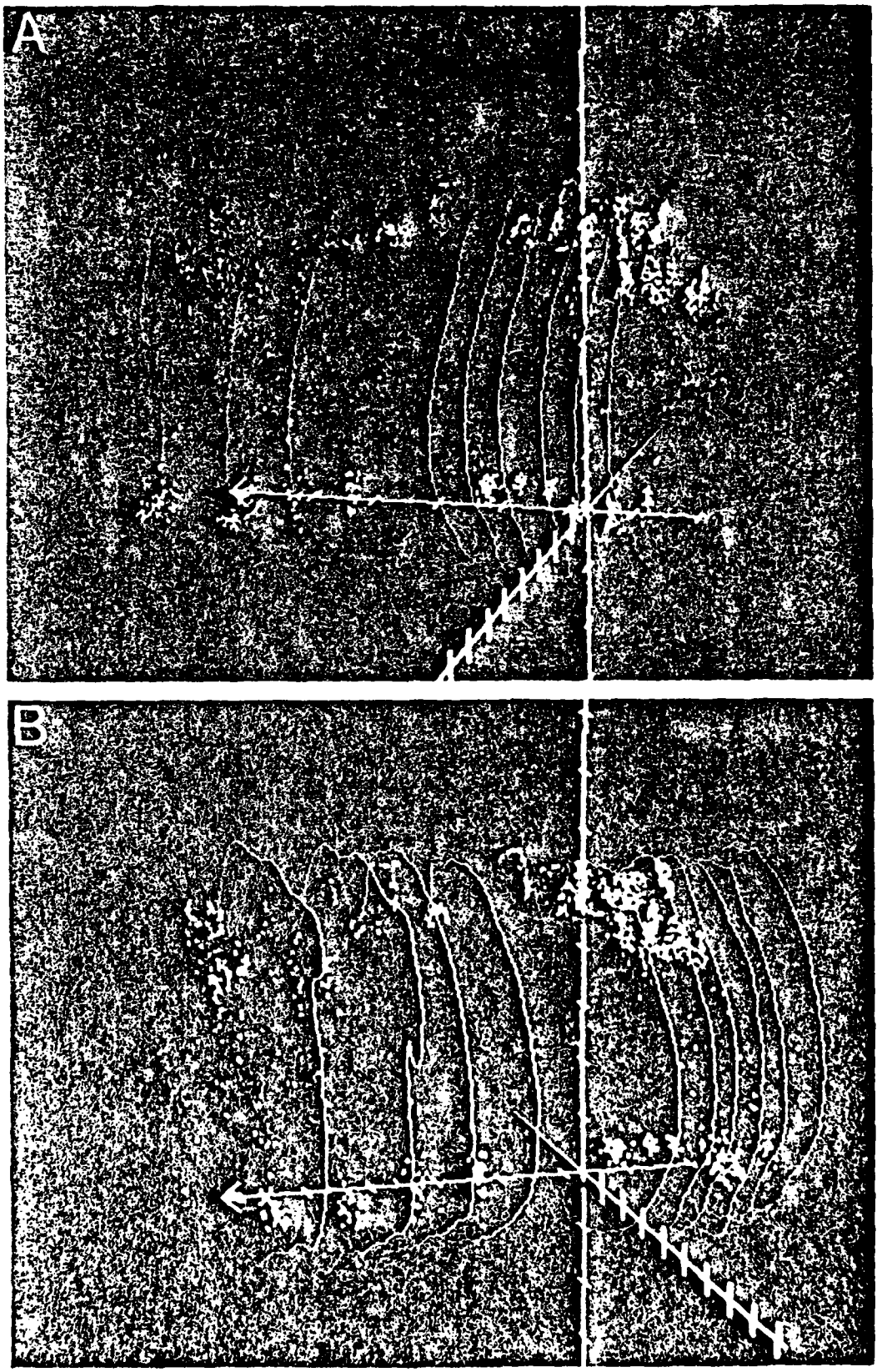


Fig 4 A Three-dimensional view of TH immunoreactive neurons in the medulla oblongata The TH positive cells are coded green The sections are viewed from the left side, with a [illegible] caudolateral tilt The caudal end appears prominently in the foreground (right of figure) B Sections shown in A containing TH immunoreactive neurons, viewed from the right side, with a 30° lateral tilt rostrally The rostral sections (left of figure) appear prominently Grid marks = 500 μm

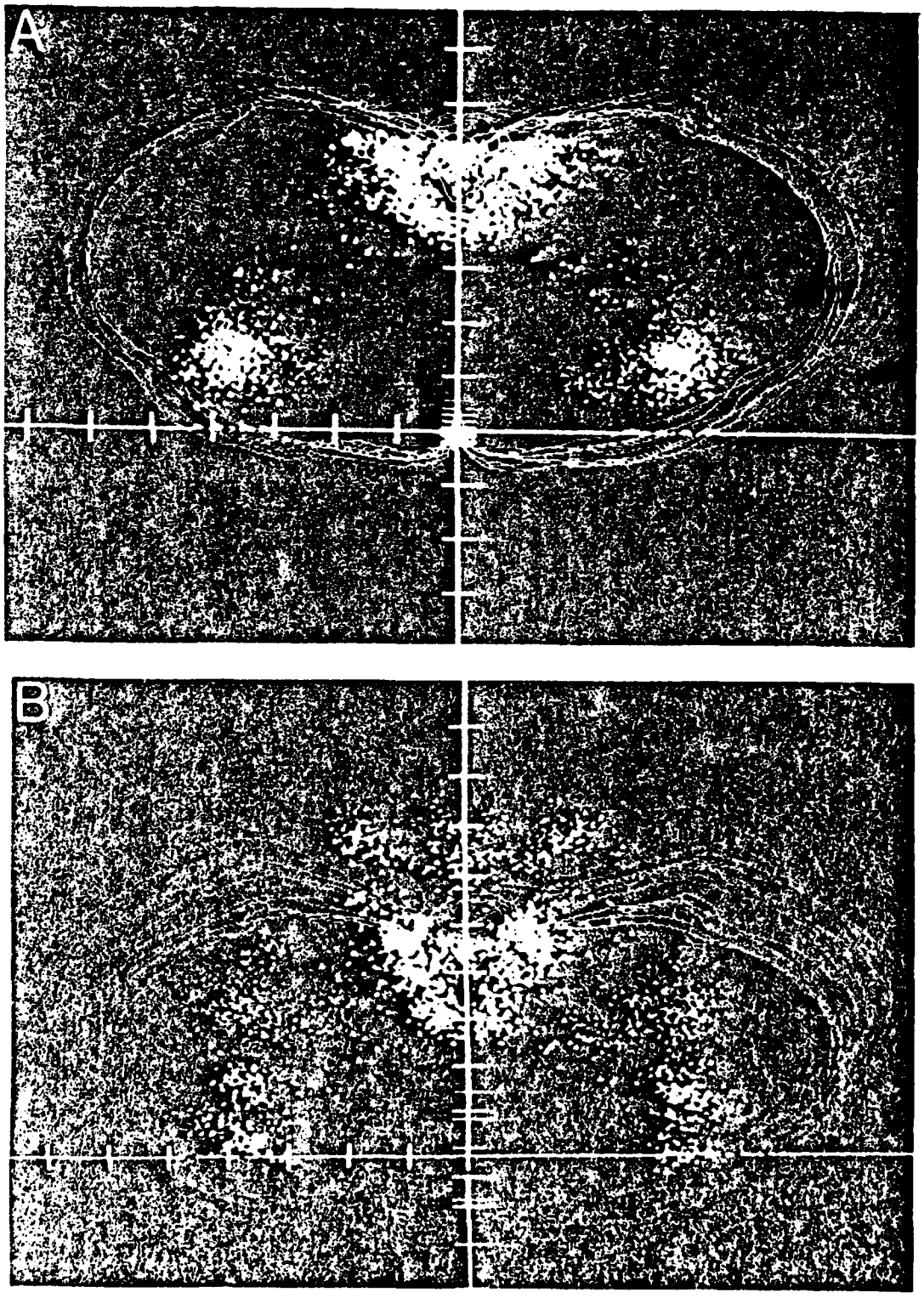


Fig 5 Three-dimensional reconstruction of TH (green), DBH (blue), and PNMT (white)-immunoreactive neurons in coronal sections of the rat medulla These sections are viewed from the caudal side with no tilt Notice the paired dorsal cell groups (A2 and C2 which appear indistinguishable from each other) and the paired ventral cell groups (A1 and C1 which appear as a single population in this reconstruction) B Thirty degree rostral tilt of the same sections as in A The caudal cell groups A2 (dorsal) and A1 (ventral) are visible in the foreground, and the rostral cell groups C2 (dorsal) and C1 (ventral) can be seen as four separate populations Grid marks = 500 μm.

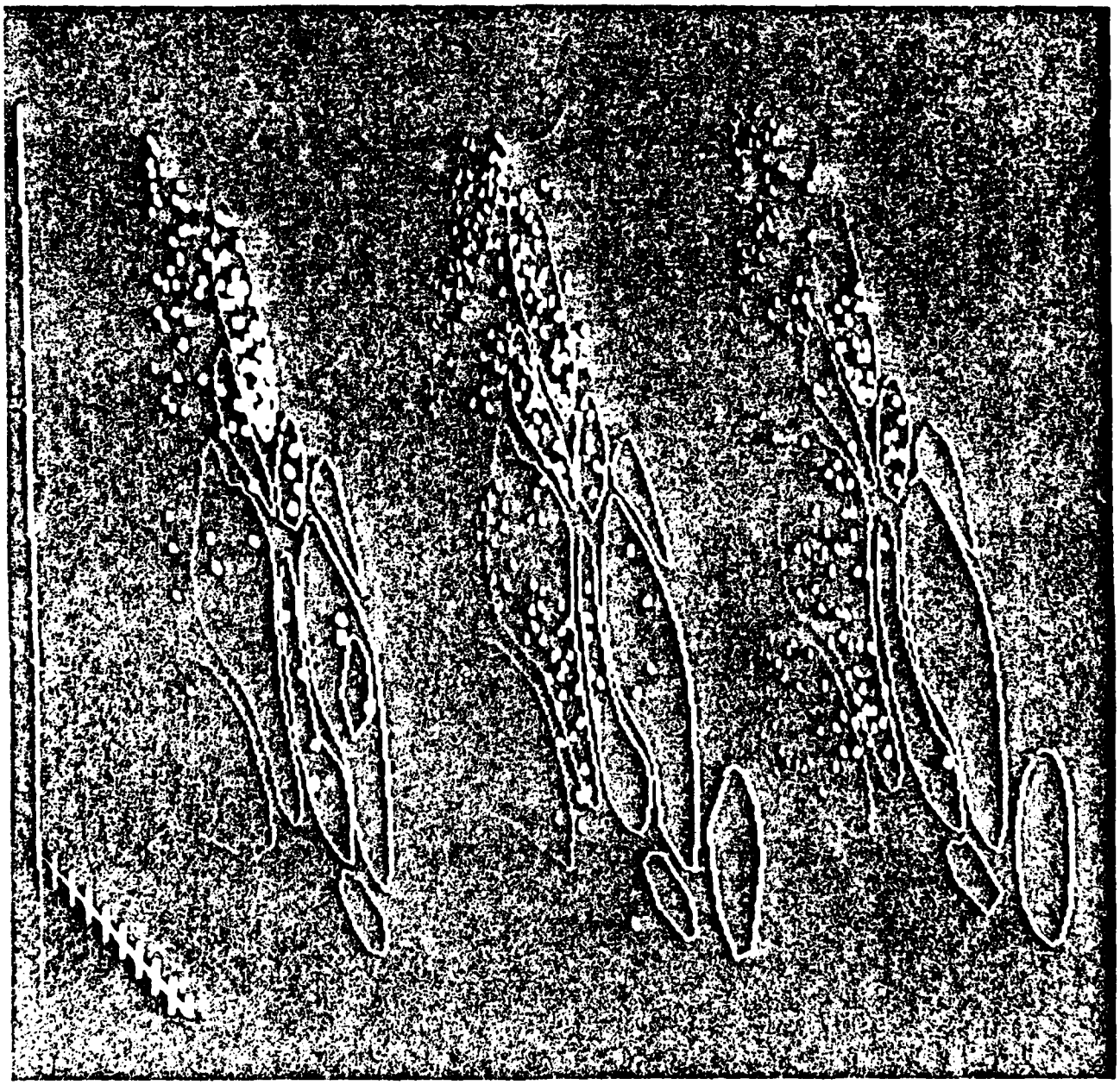


Fig. 6. Left side view of three sections of the dorsal medulla at the level of the area postrema showing, from right to left, the location of TH (green), DBH (blue), and PNMT (white) immunoreactive neurons. The sections are tilted 30° mediolaterally. The left lateral side of the sections is in the foreground, and the midline is in the background. Nuclear boundaries and levels of sections are identical with those in Fig. 2. Grid marks = 100 μm.

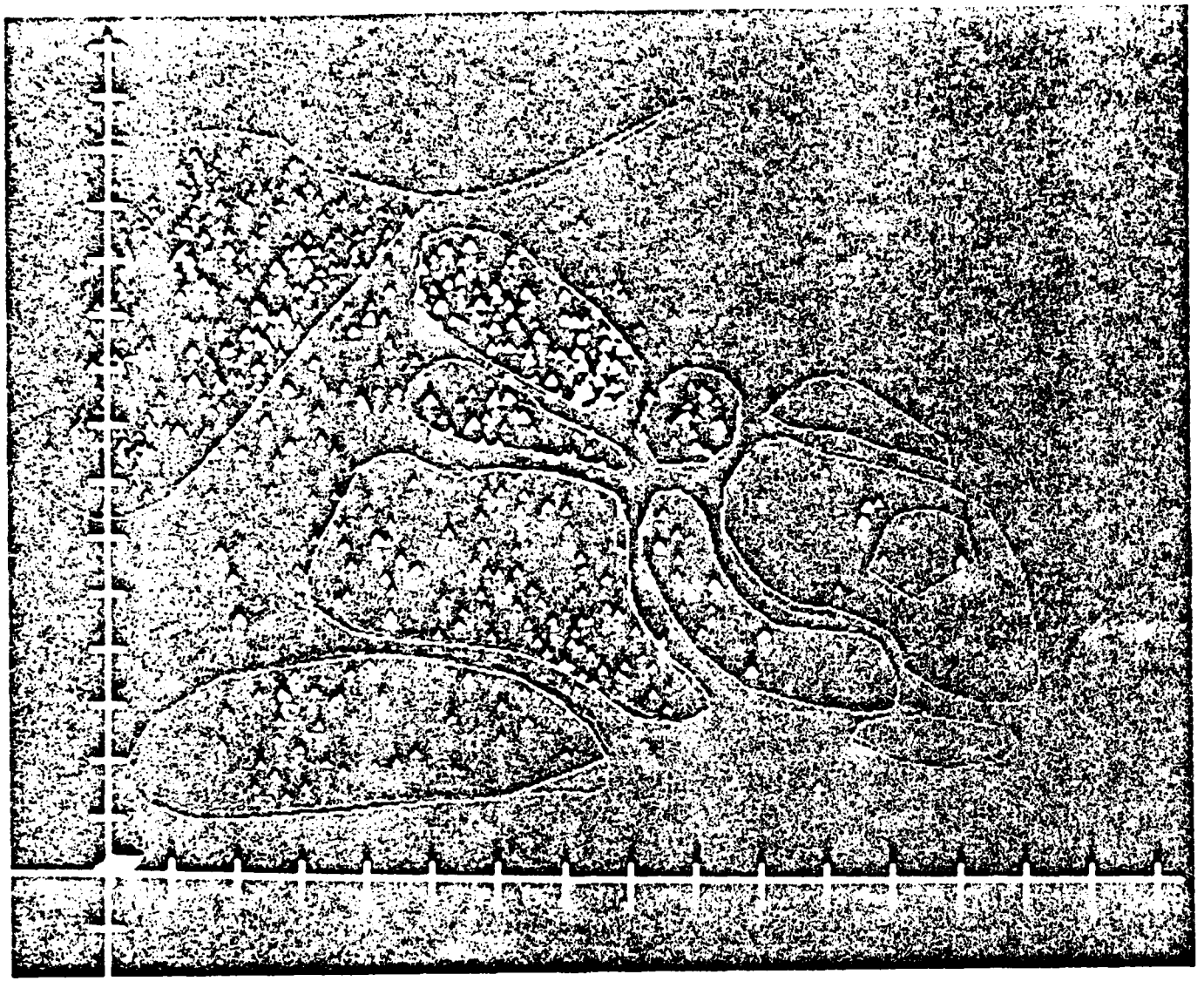


Fig 7 Frontal view of three sections of the dorsal medulla at the level of the area postrema, showing the location of TH (green), DBH (blue), and PNMT (white) immunoreactive neurons. The sections have been overlapped in order to demonstrate the correspondence between the location of cell bodies containing the three catecholamine synthesizing enzymes. Nuclear boundaries are identical with those in Figure 6. Grid marks = 100 μm.

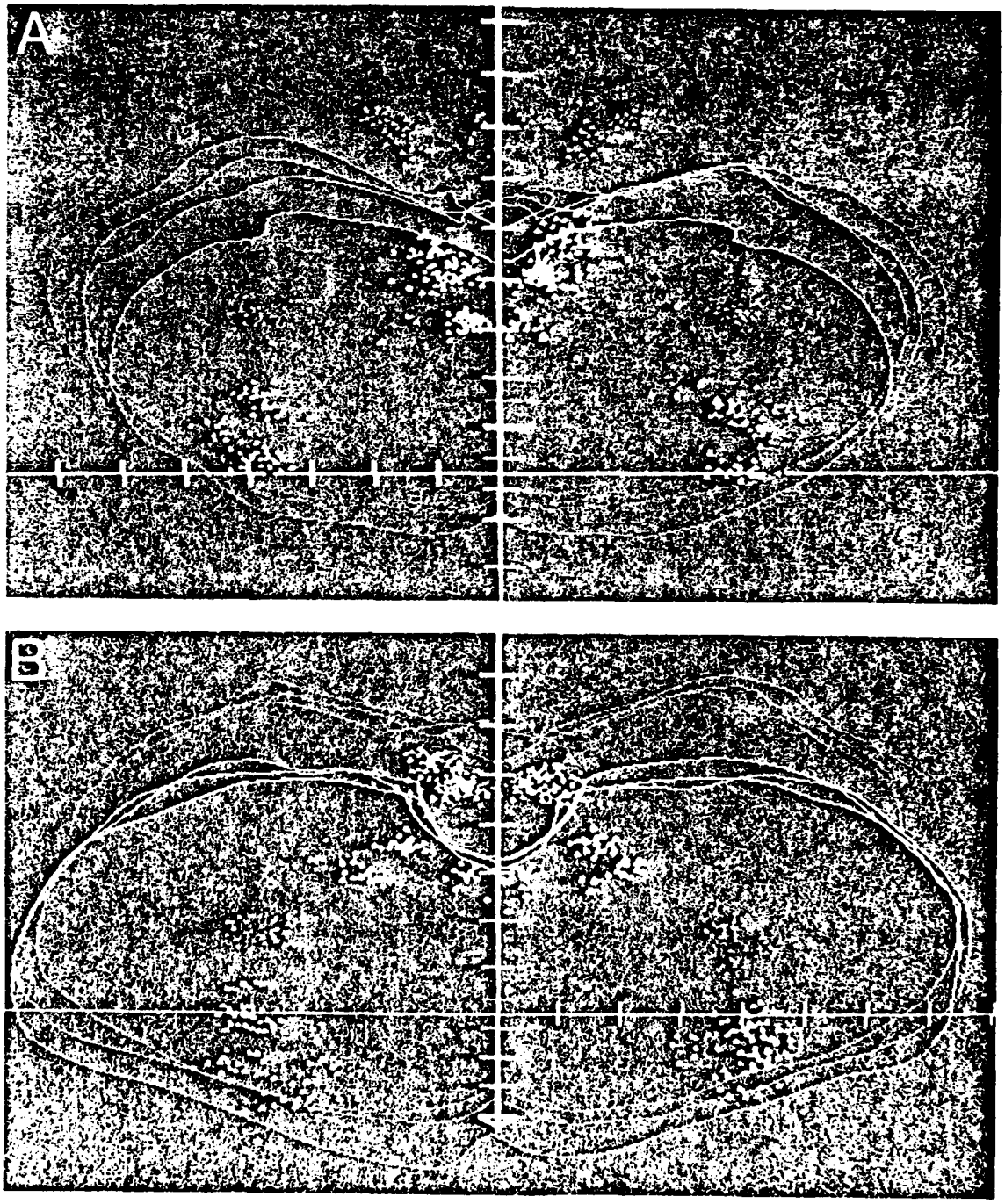


Fig 8 Three-dimensional reconstruction of DBH (blue) immunoreactive neurons in the medulla Noradrenaline and adrenaline neurons contain this catecholamine synthesizing enzyme A A view from the caudal side with a 30° tilt The A2 (dorsal) and A1 (ventral) cell groups are prominent in the foreground B. A view of the cell population shown in A from the rostral side with a 30° tilt The C2 (dorsal) and C1 (ventral) catecholaminergic cell populations can be seen in the foreground Grid marks = 500 μm

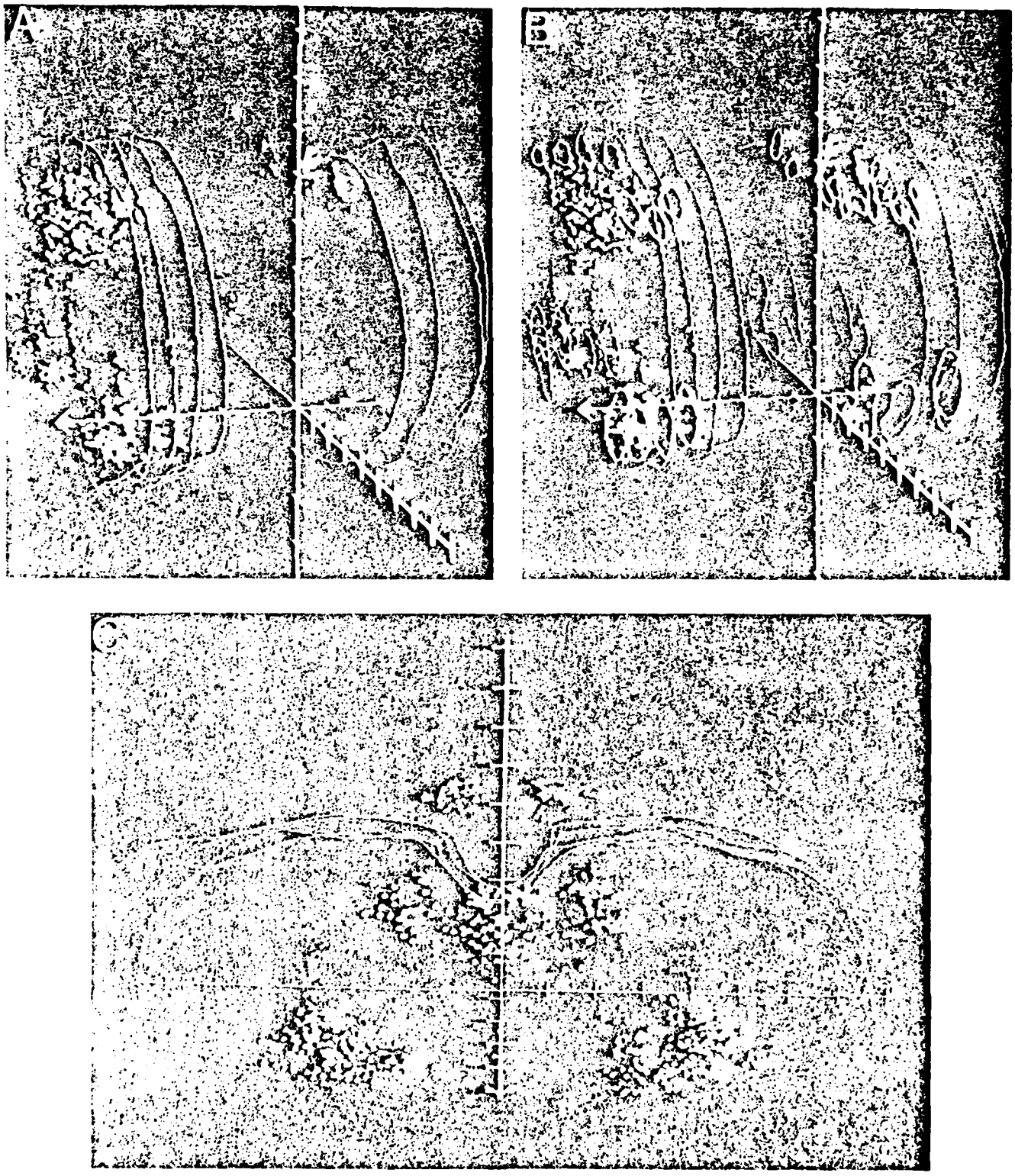


Fig. 9. A. A right lateral view of of cross sections of the medulla in three dimensions, showing the position of PNMT (white) immunoreactive neurons. This catecholamine synthesizing enzyme is only present in adrenergic neurons. The sections are tilted 30° rostrally (left of figure). The C2 (dorsal) and C1 (ventral) cell groups are prominent in the foreground (left of figure). Sections on the right (caudal levels) show the presence of adrenergic neurons in the dorsal medulla. This is a small paired population located dorsal to the tractus solitarius. B. The same reconstruction as in A, with the ventral nucleus lateral reticular nucleus caudally, and paragigantocellularis rostrally and dorsomedial tract (tractus solitarius) indicated as solid lines. Notice that the C1 cell group overlaps considerably with the paragigantocellularis (rostral sections), and the caudal sections show adrenergic neurons located dorsal to the tractus solitarius. Grid bars = 500 μm.

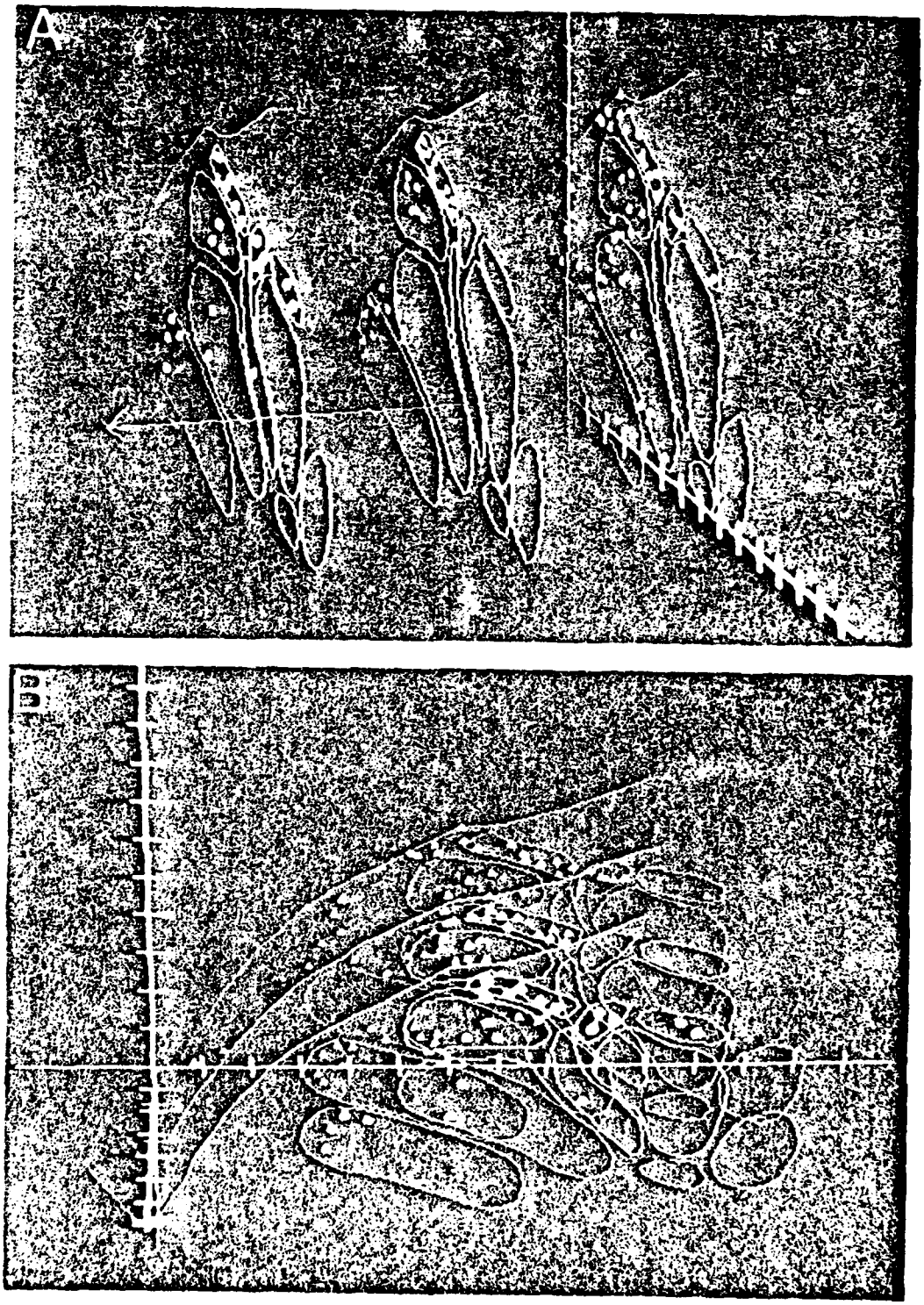


Fig 10 A High magnification sections of the *dorsal medulla* showing (from right to left) TH (green), DBH (blue), and PNMT (white) immunoreactive neurons at a level 1 mm rostral to the obex. The view is *from the* right lateral aspect, with a 30° tilt. The major nuclear boundaries are indicated by solid lines. Notice that the distribution of all three catecholamine synthesizing enzymes is similar, indicating that most of the neurons in these populations are adrenergic (TH-, DBH-, and PNMT-positive). B. Frontal view, with a 30° tilt, of the sections shown in A. Grid bars = 100 μm.

in the midline region of the area postrema. A second paired group of cells in the background is visible: the C1 (ventral) group and the C2 (dorsal) group. Figure 8B is a view from the rostral side in which the paired C1 and C2 groups can be clearly seen in the ventral and dorsal regions, respectively. The A1 cell group is a homogeneous population of NA cell extending from −2.5 to +0.2 mm with dimensions of 1.3–2.7 mm (cells staining blue only). Compare this figure with Figure 9 to determine the location of the noradrenaline cell groups. Figure 9 shows cell groups with PNMT immunoreactivity. Since neither dopaminergic nor noradrenergic neurons contain this synthesizing enzyme PNMT-containing cells can be considered to be synonymous with adrenaline containing. Figure 9A is a 30° tilted rostrolateral view of coronal sections showing the location of PNMT-immunoreactive neurons. Figure 9B is identical with 9A except that in 9B the boundaries of the lateral reticular nucleus (LRt) and the paragigantocellular reticular nucleus (PGi) on the ventral side and the tractus solitarius (TS) on the dorsal side have been plotted. The following features of adrenaline-containing neurons are obvious at first glance: (1) the majority of adrenaline cells are located rostrally; (2) they appear to be localized in the dorsomedial and ventrolateral regions but are not clustered in very distinct regions; and (3) a small paired group of adrenaline-containing neurons can be distinguished in the caudal dorsal medulla which does not appear to be a part of the more rostrally located C2 cell group. This also does not belong to the more caudally located A2 cell group. This region has been examined at high magnification in Figure 10.

Figure 10A is a three-dimensional reconstruction of three consecutive sections showing (from left to right) TH-, DBH-, and PNMT-positive immunoreactive cells, respectively. All three sections show identical locations of immunoreactive neurons indicating that this region is predominantly adrenergic. The location of adrenaline-containing cells in the dorsal strip (ds) region can be clearly seen.

DISCUSSION

The results of this study demonstrate that catecholaminergic neurons are organized according to cytoarchitectonic boundaries and rostrocaudal location in the medulla oblongata. The A2 cell group, previously considered to be a homogeneous noradrenergic cell group, constitutes a heterogeneous population of neurons containing a variety of cell types and all three types of catecholaminergic neurons. In this section we shall first discuss the strengths and weaknesses of the three-dimensional computer reconstruction technique which was used to analyze this complex data. In the second part of this discussion we shall examine the significance of this new data in light of our understanding of medullary catecholaminergic systems.

Three-dimensional computer reconstruction

Use of three-dimensional computer reconstruction has enabled us to view the catecholaminergic cell groups in the medulla oblongata in a unique manner. The ability to manipulate the data in a variety of different ways makes possible an accurate analysis of neuroanatomical systems in a way that physically cannot be done by artists. The use of color coding in this series has further permitted us to study multiple types of data simultaneously in the same section with or without representation of nuclear boundaries. In these studies, the task of determining what pattern of immunocytochemically identified cell groups corresponds to what specific neurotransmitter system is facilitated by the use of color computer graphs. Computer reconstruction of the distribution patterns of transmitter-identified neurons aids in the conceptual analysis of this system.

In order to appreciate the view or the concept of the separation or conversely, the overlap or the rostrocaudal extent of the cell group from a series of classical neuroanatomical drawings, there is a need to examine each of these drawings simultaneously to construct a mental image of where the cell groups are located. The combined data bases from different sections and the different views subsequently synthesized by means of the computer eliminate the need to section the brain in all three planes to obtain different visualizations.

Computer-generated rotation of the orientation of the sections allows one to separate out different cell groups and permits the visualization of functionally distinct and immunocytochemically distinct neuronal populations (Figs. 8, 9). Figure 5 also shows that the computerized rotation and tilting of the serial reconstruction of cells is an extremely valuable tool in the separation and identification of different populations of neurons. This method requires only a few seconds of viewing for the entire picture to be integrated by the human brain and offers a great advantage over classical neuroanatomical drawings. The expense of publishing color plates can be considered to be a major disadvantage of this method of analysis. However, the cost of eight consecutive pages (one signature) in *The Journal of Comparative Neurology* does not significantly exceed the cost of a single color plate. The savings in medical illustrator charges alone offsets this cost, thus making the publication of this data cost effective and affordable to most neuroanatomical laboratories provided the computer system itself is shared between many users.

The nucleus of the tractus solitarius and the related regions of the medulla oblongata provide an ideal anatomical location; where a large number of monoaminergic nerve cells are distributed, use of computer reconstruction may enhance conceptual analysis as compared to classical neuroanatomical drawings. In the two preceding papers (Kalia et al., '85a,b) we prepared a series of very accurate drawings. These anatomical maps were stored in the computer and later could be retrieved, rotated, tilted, viewed from different angles, overlapped, or viewed individually. Thus multiple drawings of different neuronal populations of cells were plotted.

In the present paper the computer aided the visualization task by estimating where a population of transmitter-identified neurons were present. This was based on visual overlap of different groups of immunoreactive neurons.

In the medulla oblongata there exist three populations of neurons: dopaminergic, noradrenergic, and adrenergic neurons. Immunocytochemistry of this system does not directly provide us with information about the neurochemical identity of these catecholaminergic neurons. Rather, we get information about the presence or absence of the catecholamine-synthesizing enzymes from which we have to deduce the existence of the neurotransmitter. This system, therefore, requires a number of steps in the analysis, most of which are based on a simultaneous visualization and synthesis of this data. The individual position of neurons showing immunoreactivity to the three catecholamine-synthesizing enzymes (TH, DBH, and PNMT) was defined in three different planes by the computer, enabling us to analyze the data in a variety of ways.

The overall organization and heterogeneity of medullary catecholaminergic neurons

Quantitation of catecholaminergic cell bodies as well as three-dimensional reconstruction with stereotaxic coordinates as shown in this study have helped answer a number of questions. Where in the three-dimensional space are adrenaline neurons in the nTS located? Where in the three-

dimensional space are dopamine and adrenergic neurons located? Do the C1 and the A1 cell groups overlap, and if they do what is the rostrocaudal level at which they overlap? Do the A1 and A2 cell groups overlap, and if so what is the level of the overlap? What is the extent of the overlap and do these cell populations merge and continue at the same level in the mediolateral plane or do they diverge? What is their relationship to other nuclear groups? These questions have been answered by the analysis presented in this paper.

The present results demonstrate that the medullary catecholaminergic neurons are organized in a complex manner that is not based either on cytoarchitecture or rostrocaudal location. The homogeneous neurochemically distinct cell populations of cells such as the A1 (noradrenergic) and C1 (adrenergic) appear to consist of subpopulations of neurons which cannot be separated functionally on the basis of cytoarchitectonics alone. Clearly, connectivity studies need to be done to study these systems in greater detail. The heterogeneous catecholaminergic cell group (A2) in the caudal, dorsal medulla, on the other hand, is of considerable interest since this region is known to contain functionally distinct subpopulations of neurons (Kalia and Mesulam, '80).

Thus the observations regarding specific dopaminergic neurons in the mid-line of the area postrema, the periventricular region, and the dorsal motor nucleus of the vagus might be of considerable physiological significance since these regions are known to be related to the emetic system of the medulla.

The noradrenergic neurons in the A2 cell group have been known to exist since 1964 (Dahlström and Fuxe). However, the detailed pattern of distribution of these noradrenergic neurons in the various subnuclei of the nucleus of the tractus solitarius and adjacent regions of the dorsal medulla indicate that this noradrenergic system is involved in multiple visceral functional effects (Kalia and Mesulam, '80).

The finding that adrenergic neurons (TH, DBH, and PNMT positive) are located in the A2 cell group is new. The clustering of these adrenergic neurons in the dorsal strip region and the dorsal subnucleus of the nTS indicates that these neurons might be related to cardiovascular effects since carotid sinus nerve and aortic nerve afferents are known to terminate in this region (Kalia and Welles, '80).

Finally the adrenergic, the C2 cell group, which was originally described by Hökfelt et al. ('74) as being a "comparatively small group," is in fact a large cell population (Fig. 9) scattered over a large region of the medulla, covering a number of nuclei in the rostral part of the dorsal medulla: the medial longitudinal fasciculus, the dorsal motor nucleus of the vagus, and the prepositus hypoglossal nucleus. The morphological characteristics of this cell group are very different from those of the A2 cell group and thus cannot be considered to be a rostral extension of that cell population. In addition the analysis in this paper revealed a distinct separation between the C2 and A2 cell groups.

In summary, the striking features of three-dimensional computer reconstruction of the location of catecholamine-synthesizing enzymes in the medulla oblongata reveal the existence of a number of new features of medullary catecholaminergic neurons. The morphological complexity, detailed organization, and precise localization of this system of neurochemically distinct neurons indicates their functional heterogeneity. This data must be considered in the context of the patterns of connectivity of this region of the brain stem.

ACKNOWLEDGMENTS

This work was supported by USPHS grants HL 30991, HL 33632, and HL 31997, and American Heart Association grant 81-978 to M.K., USPHS grant NIAAA 390 and DA 2338, and a grant from the Biol. Humanics Foundation to D.J.W. and USPHS grant MH 25504 and Swedish Medical Research Council grant 14X-04246-10B to K.F. It is a pleasure to thank Kaveri Kalia for secretarial and technical assistance. We are grateful to Marc Lebowitz, Mike Leddy, Paulie Tartaglia, Mark Brooks, and Karen Gazzard for their expert assistance with the photomicrographs.

LITERATURE CITED

Borison, H.L., and S.C. Wang (1949) Functional localization of central coordinating mechanisms for emesis in the cat. J. Neurophysiol. *12* 305-313.

Dahlström, A., and K. Fuxe (1964) Evidence for the existence of monoamine containing neurons in the central nervous system. I. Demonstration of monoamines in the cell bodies of brainstem neurons. Acta Physiol. Scand. *62/Suppl. 232* 1-55.

Falck, B., and N.-A. Hillarp (1962) On the cellular localization of catecholamines in brain. Acta Anat. *38* 277-279.

German, D.C., D.S. Schlusselberg, B.A. McMillen, K. McDermott, W.K. Smith, and D.J. Woodward (1982) Asymmetries in human brain dopamine receptor binding: Relationship to 3-dimensional reconstruction of midbrain dopamine neurons. Neurosci. Abstr. *8* 303.

German, D.C., K.L. McDermott, M.K. Sanghera, D.S. Schlusselberg, W.K. Smith, D.J. Woodward, S.G. Speciale, and C.B. Saper (1983) Three-dimensional reconstruction of dopamine neurons in the mouse: Strain differences in regional cell densities and pharmacology. Neurosci. Abstr. *9* 1150.

Goldstein, M. (1972) Enzymes involved in the catalysis of catecholamine biosynthesis. In R.N. Ubell (ed) Methods in Neurochemistry. Vol. 1, New York: Plenum Press, pp. 317-340.

Hökfelt, T., K. Fuxe, M. Goldstein, and O. Johansson (1973) Evidence for adrenaline neurons in the rat brain. Acta Physiol. Scand. *89* 286-288.

Hokfelt, T., K. Fuxe, M. Goldstein, and O. Johansson (1974) Immunohistochemical evidence for the existence of adrenaline neurons in the rat brain. Brain Res. *66* 235-251.

Kalia, M., K. Fuxe, and M. Goldstein (1985a) Rat medulla oblongata. II. Noradrenergic neurons, nerve fibers and preterminal processes. J. Comp. Neurol. *233* 308-332.

Kalia, M., K. Fuxe, and M. Goldstein (1985b) Rat medulla oblongata. III. Adrenergic neurons, nerve fibers and preterminal processes. J. Comp. Neurol. *233* 333-349.

Kalia, M., K. Fuxe, T. Hokfelt, O. Johansson, R. Lang, D. Ganten, C. Cuello, and L. Terenius (1984) Distribution of neuropeptide immunoreactive nerve terminals within the subnuclei of the nucleus of the tractus solitarius of the rat. J. Comp. Neurol. *222* 409-444.

Kalia, M., and M.M. Mesulam (1980) Brain stem projections of sensory and motor components of the vagus nerve in the cat. II. Laryngeal, tracheobronchial, cardiac and gastrointestinal branches. J. Comp. Neurol. *193* 467-508.

Kalia, M., and R.V. Welles (1980) Brain stem projections of the aortic nerve in the cat: A study using tetramethyl benzidine as the substrate for horseradish peroxidase. Brain Res. *188* 23-32.

Reis, D.J., R.H. Benno, L.W. Tucker, and T.H. Joh (1982) Quantitative immunocytochemistry of tyrosine hydroxylase in brain. In V. Chan-Palay and S. Palay (eds) Cytochemical Methods in Neuroanatomy. New York: Alan R. Liss Inc., pp. 205-225.

Schlusselberg, D.S., W.K. Smith, B.G. Culter, and D.J. Woodward (1982a) A computer system for semi-automatic cell recognition in neuroanatomic studies. Neurosci. Abstr. *8* 644.

Schlusselberg, D.S., W.K. Smith, M.H. Lewis, B.G. Culter, and D.J. Woodward (1982b) A general system for computer based acquisition, analysis and display of medical image data. Proc. ACM 18-25.

Smith, W.K., D.S. Schlusselberg, and D.J. Woodward (1981) A computer system for neuroanatomical data acquisition, analysis, and display. Neurosci. Abstr. *7* 135,18.

Sternberger, L.A. (1979) Immunocytochemistry. New York: J. Wiley.

Reprint Series
15 March 1985, Volume 227, pp. 1351–1354

SCIENCE

A Renal Countercurrent System in Marine Elasmobranch Fish: A Computer-Assisted Reconstruction

Eric R. Lacy, Enrico Reale, Daniel S. Schlusselberg, Wade K. Smith, and Donald J. Woodward

A Renal Countercurrent System in Marine Elasmobranch Fish: A Computer-Assisted Reconstruction

Abstract. *Computer-aided techniques were used to reconstruct the complex renal tubular system in the dorsal kidney region of a marine elasmobranch fish, the little skate (*Raja erinacea*), from a series of light micrographs of serial sections. It was established that five individual segments of one nephron, consisting of two loops and a distal tubule, are arranged in parallel within an elongated closed tissue sac. Capillaries, which form a network around these nephron segments, enter and exit this sac at the same end. This anatomical arrangement suggests that a complex renal countercurrent multiplier system may be important in fluid regulation in these fish.*

Typically, marine fish are faced with the problem of dehydration due to the high osmolality of the surrounding seawater (*1*). Marine elasmobranch fish appear to resolve the problem by maintaining high urea concentrations in plasma and tissue, thereby elevating the osmolality of their internal milieu to nearly that of the surrounding seawater (*2*). An extremely small fraction of urea is excreted (*3*), and studies of the kidney have been done to determine how this urea conservation is achieved (*1–5*). However, the extremely complex configuration of the elasmobranch nephron has impeded physiological studies of the anatomical site of urea reabsorption and of some of the cellular mechanisms that are involved (*6*). This nephron complexity is evidenced by the fact that elasmobranchs are members of the only vertebrate class (Chondrichthyes) in which the nephron configuration and epithelial sequence along its length to the collecting duct are not known (*6, 7*).

We used computer graphics to study nephron configuration. A three-dimensional reconstruction of the tubule in the complex dorsal region of a marine elasmobranch kidney shows that parallel tubular segments from a single nephron are tightly wrapped in a cellular sheath that also encloses a capillary network around the tubules. This anatomical arrangement presumably has the potential of facilitating a physiological countercurrent system.

The kidneys of little skates (*Raja erinacea*) and spiny dogfish (*Squalus acanthias*) were fixed with buffered glutaraldehyde by vascular perfusion (*8*). We used 118 light micrographs of serial sections from the skate to outline and sequence 2945 tubule contours (outer circumference of tubules) of the sectioned nephron as it completed its course within the dorsal region of the kidney. The same photographs were used to outline the cellular peritubular sheath that surrounds bundles of these tubule segments (*9*). Because a complex computer-generated image was expected, the profiles of the blood vessels accompanying the tubules were not entered into the computer. The course of these vessels was simpler than that of the adjacent tubules and it was traced in the same serial sections by light microscopy.

The architectural complexity of nephrons in the dorsal region was greater than what could be reconstructed by techniques such as wax modeling. For the computer-assisted reconstruction, we used a hierarchical data-base design that accommodates the complex branching nature of biological structures and performs coordinate transformations necessary to convert serially sectioned biological materials into three-dimensional display coordinates. Solid modeling algorithms were used to generate the video images (*10, 11*).

Beginning with the urinary pole of the renal corpuscle, we traced the course of the nephron as it first entered the bundle zone in the dorsal region of the kidney

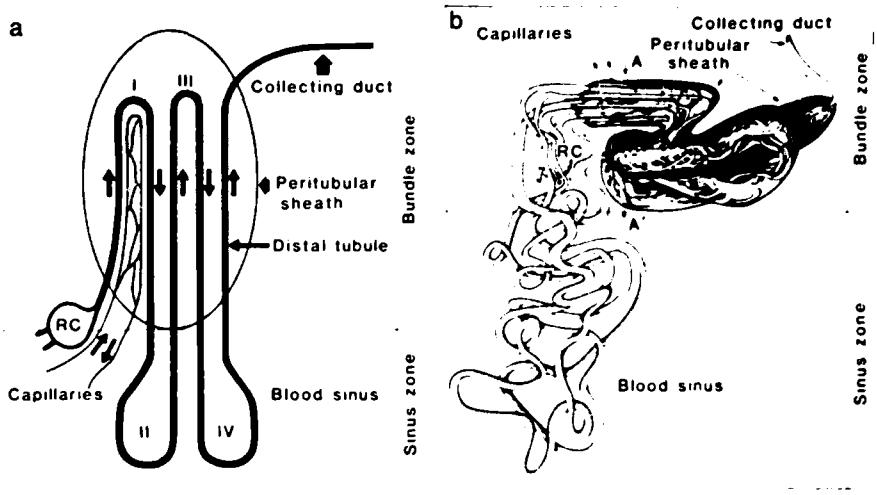


Fig. 1. Diagrams of elasmobranch nephron in the bundle zone (dorsal) and sinus zone (ventral). The dorsal kidney surface is parallel to the top of the page. (a) Simplified diagram showing renal corpuscle (RC) and four highly stylized nephron loops (I to IV). A peritubular sheath surrounds the countercurrent system of nephron segments (loops I, III, and the distal tubule) and anastomosing capillary loops in the bundle zone. Small arrows indicate the direction of tubular fluid and blood flow. (b) Schematic drawing of the pathway of the skate nephron in the bundle zone (dorsal) and in the sinus zone (ventral) showing some of the nephron complexity. The entering limbs of nephron loops I and III and the distal tubule (a) pierce the peritubular sheath near the renal corpuscle and extend to the opposite end of the sheath. Close to the renal corpuscle, the five tubular segments (loops I, III, and the distal tubule) located in the bundle zone are covered by the peritubular sheath and run parallel to each other. To emphasize this distinctive course, they have been drawn side by side in one plane and not assembled into a bundle, as they actually are (Figs. 2 and 3). The tubular bundle and surrounding peritubular sheath then become convoluted. The parallel course of the tubules is lost since the loops wrap around each other. For simplicity, the opposite end of the peritubular sheath, where the distal tubule emerges, has been drawn away from the renal corpuscle on the far right side of the diagram. The distal tubule pierces the sheath at this point to join a collecting duct, whereas the two other nephron segments loop back and retrace their path, finally exiting the sheath where they entered it. Capillaries also enter and exit the peritubular sheath at its renal corpuscle terminus and form an anastomotic network around and within the tubular bundle. A histological section perpendicular to the plane of the drawing and along the line A–A' is shown in Fig. 2. In the sinus zone, loops II and IV meander in large blood sinuses.

and made the first of four large loops (Fig. 1). After forming a first dorsal loop, the tubule turned and exited the bundle zone to reach the sinus zone in the ventral part of the kidney. There the nephron formed a second loop (loop II in Fig. 1), returned to the bundle zone, and with a third loop (loop III in Fig. 1) returned to the sinus zone. Finally, after a fourth loop (loop IV in Fig. 1), the nephron reentered the bundle zone and, as the distal tubule, it joined a collecting duct in the subcapsular region. In the bundle zone, both loops I and III (each composed of two limbs, one ascending and one descending) and the early distal tubule are close together (Fig. 1a). This bundle of five tubular segments, all from the same nephron, was wrapped by a cellular peritubular sheath separating them from other peritubular sheaths (each containing two nephron loops and an early distal tubule) derived from other nephrons (Figs. 1 and 2). The peritubular sheath (Figs. 1 and 2) was observed to be an elongated closed sac formed by sever-

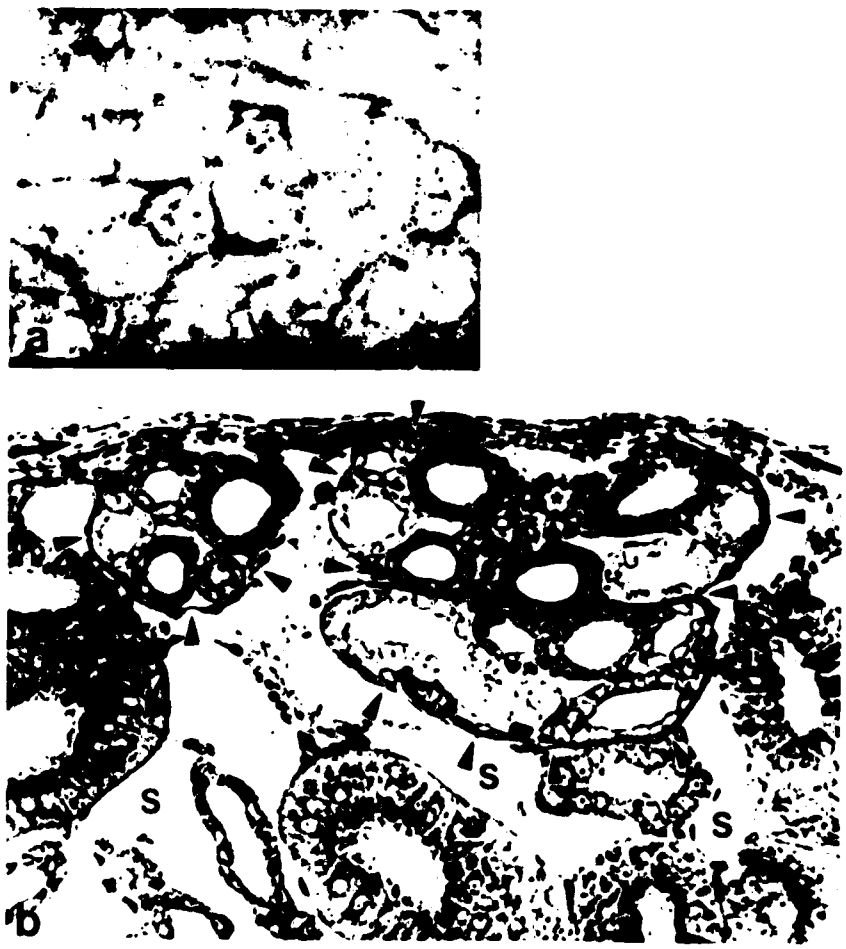


Fig. 2. (a) Photomicrograph of dorsal skate kidney surface showing a field of numerous tightly intertwined nephron segments. Two discrete tubular bundles are outlined, with nephrons visible through the peritubular sheath. The tubular bundle here corresponds to the reconstructed tubular bundle shown in Fig. 3d. (b) Photomicrograph of epoxy resin section cut through the dorsal region of skate kidney in a plane similar to that along A–A' in Fig. 1b. Three regions of the same tubular bundle and surrounding peritubular sheath (arrowheads) in the bundle zone are shown. On the upper left are five cross-sectional nephron segments and capillaries (small white areas between tubules) of the straight part of the countercurrent system. On the upper right, the tubules in the peritubular sheath are looping back so that those on the right are continuous with those on the left (star). Just below this, on the right, the sheath (arrowheads) encloses capillaries and the five tubules cut in a convoluted part of the tubule bundle. S, individual tubules in the sinus zone. Arrows indicate the dorsal kidney capsule. Magnification, ×200.

al investing layers of squamous cells that surround the tubules. Freeze-fracture and thin sections have shown tight junctions between adjacent squamous cells of the peritubular sheath. This peritubular sheath was penetrated on one end, near the renal corpuscle, by entering and exiting tubule segments of the two nephron loops and distal tubule and by capillaries that formed an anastomosing network around these five tubule segments (Fig. 1). At the opposite end, the peritubular sheath was pierced only by the exiting distal tubule. In contrast, the segments of the nephron in the sinus zone of the kidney were less organized and were not surrounded by a peritubular sheath but instead intermingled randomly with other nephron segments in large blood sinuses (Figs. 1 and 2).

In our reconstructed images of the skate nephron (Fig. 3), this bundle of five tubules (ascending and descending limbs of nephron loops I, III, and the early distal tubule as diagrammed in Fig. 1a) and the surrounding peritubular sheath (Fig. 3) continued in a straight course from near the renal corpuscle just under the skate kidney surface for approximately 0.25 mm (285 μm). The tubules within the straight part of the peritubular sheath were arranged in a parallel fashion. The peritubular sheath then became convoluted, twisting back under the straight portion of the sheath and making several additional turns. The convoluted part of the peritubular sheath was slightly more than 3 mm long (3163 μm) in the skate, as determined from the reconstructed path length. The nephron segments within the convoluted part of the sheath also became somewhat convoluted as they wound around each other, and the strict parallelism observed in the straight part of the sheath was not maintained (Fig. 3). The total path length of nephron loops I, III, and the early distal tubule were 4902, 8247, and 3636 μm, respectively (*12*). Capillaries observed in the histological sections entered and exited the end of the peritubular sheath along with the tubules, as shown in Fig. 1b. The capillaries formed an anastomotic network around and within the tubular

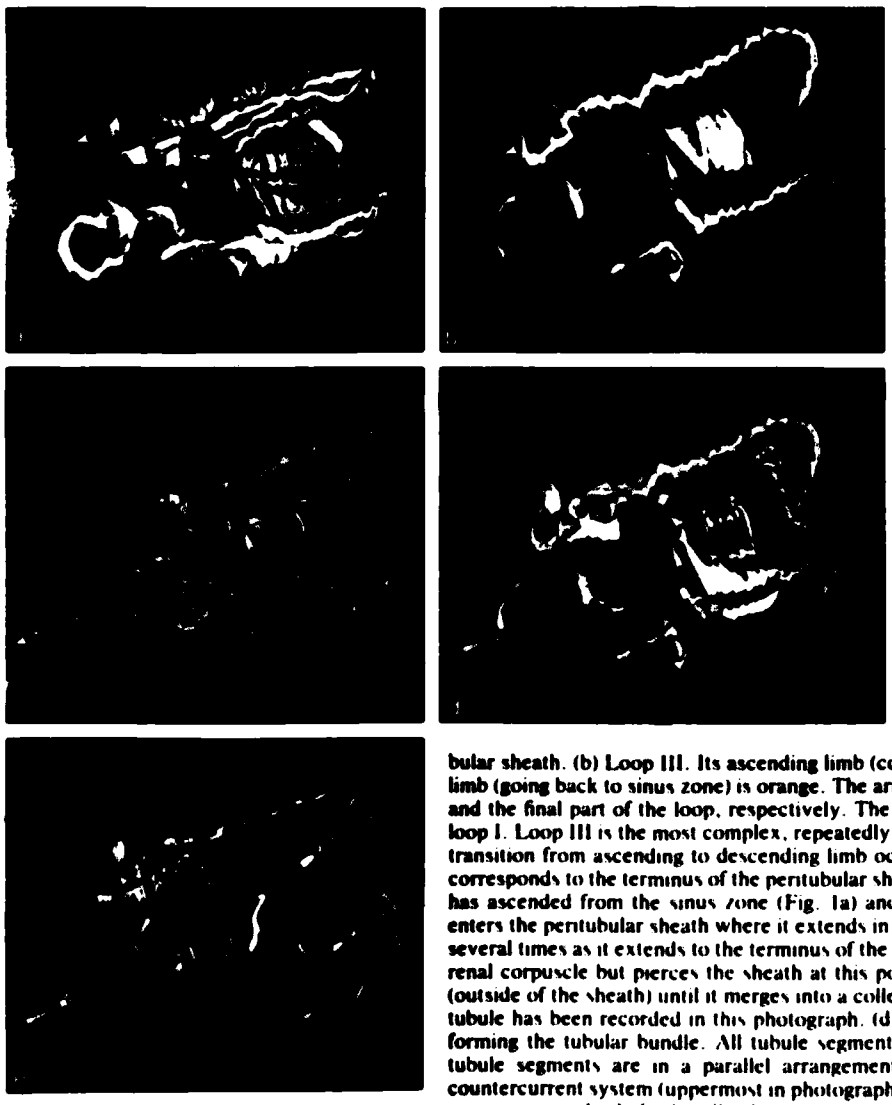


Fig. 3. Computer reconstruction of skate tubular bundle in the bundle zone (dorsal region) and its peritubular sheath. All segments shown are part of the same nephron. The dark blue calotte-shaped structure is part of the renal corpuscle close to the urinary pole. For better visualization of the tubules, the remainder of the renal corpuscle [dashed line in (a)] has not been reconstructed here. The straight part of the countercurrent system is uppermost in each photograph, as is the dorsal kidney surface. In these reconstructions the terminus of the peritubular sheath is in the lower left corner of each photograph. (a) Loop I (see Fig. 1a). The ascending (entering) limb is light green and the descending (exiting) limb is dark green. The arrow is in the ascending limb, close to its starting point from the urinary pole of the renal corpuscle, which is hidden in this perspective. After emerging from the renal corpuscle, the tubule goes left and then turns sharply right and runs along the straight part of the countercurrent system. At the upper right, the ascending limb bends back and after numerous convolutions turns to become the descending limb. The transition from ascending to descending limb occurs at the lower left of photograph. The limbs are tightly wrapped around each other as the descending limb retraces the path of the ascending limb. The arrowhead is in the final part of the loop after it has exited the peritubular sheath. (b) Loop III. Its ascending limb (coming from sinus zone) is red and its descending limb (going back to sinus zone) is orange. The arrow and arrowhead are in the beginning segment and the final part of the loop, respectively. The basic pattern of this loop is the same as that of loop I. Loop III is the most complex, repeatedly turning and wrapping itself tightly together. The transition from ascending to descending limb occurs at the lower left of the photograph, which corresponds to the terminus of the peritubular sheath (d). (c) Distal tubule. This nephron segment has ascended from the sinus zone (Fig. 1a) and starts close to the renal corpuscle (arrow). It enters the peritubular sheath where it extends in a straight path as the early distal tubule. It turns several times as it extends to the terminus of the bundle (arrowhead). It does not loop back to the renal corpuscle but pierces the sheath at this point, where it continues as the late distal tubule (outside of the sheath) until it merges into a collecting duct. Only the initial part of the late distal tubule has been recorded in this photograph. (d) Composite of loops I, II, and the distal tubule forming the tubular bundle. All tubule segments have the same positions as in (a) to (c). The tubule segments are in a parallel arrangement along the length of the straight part of the countercurrent system (uppermost in photograph). Each outlined bundle of Fig 2a corresponds to a reconstructed tubular bundle shown here. (e) The peritubular sheath surrounding the tubular bundle. The renal corpuscle has been removed and is outlined by dashes. The perspective is the same as in (a) to (d). The closed end of the sheath near the renal corpuscle is pierced by the five tubular segments: entering and exiting limbs of loops I (light green and dark green) and III (red and orange) and the entering distal tubule (blue), where the straight part of the countercurrent system begins. The early distal tubule (inside the sheath) exits the sheath at its terminus in the lower left corner of the photograph, where it becomes the late distal tubule. After the straight part of the countercurrent system (upper right) the numerous turns of the tubular bundle are more easily visualized by the large sweeping turns of the peritubular sheath. Capillaries are not reconstructed here but would enter and exit the peritubular sheath near the renal corpuscle as shown in Fig. 1b.

bundle as they extended to the end of the sheath. Serial sections of the shark kidney, although not used for computer graphics, showed tubule bundles (each wrapped by a peritubular sheath) composed of linearly arranged nephron segments and capillaries like those in the skate.

This parallel arrangement of tubules and capillaries has the potential to operate as a countercurrent multiplier system. The complete countercurrent flow system consists of a straight bundle of five tightly packed tubules from a single nephron, three of which pass in one direction and two of which travel counter to that direction. Among the tubules there is a system of anastomosing capillaries that also run the length of the tubular bundle. Both tubules and capillaries are encased by the peritubular sheath.

Tubular fluid from the glomerular ultrafiltrate travels in the same direction in the ascending limb of the first loop (I) as tubular fluid in the ascending limb of the third loop (III) and the tubular fluid in the distal tubule (Fig. 1). Countercurrent flow occurs in the descending limbs of both dorsal loops (I and III). The dual character of the system, in which the two countercurrent flows are parallel to one another, is unusual. Presumably fluid within the space contained by the peritubular sheath is shared by the exterior of the five enclosed tubules and capillaries.

Similarities to mammalian kidney structure suggest that marine elasmobranch fish may employ renal countercurrent mechanisms to achieve osmotic balance. The mammalian kidney is arranged so that descending and ascending loops of Henle and collecting ducts share extracellular fluid in the renal medulla, which contains a network of descending and ascending capillaries. In these fish, the peritubular sheath probably seals each bundle of five tubules and associated capillaries from the rest of the organ, thus forming a discrete anatomical unit. This may create some functional similarities to the conditions within the mammalian renal medulla, in which there is a net reabsorption and subsequent recycling of urea (*13*). However, the renal countercurrent multiplier system in mammals, which has the capacity to produce a hypertonic urine as well as to reabsorb most urea, must be significantly different from that of the elasmobranch kidney, which conserves urea but does not produce an osmotically concentrated urine (*1, 2*). Furthermore, finding the anatomical requisites for a renal countercurrent multiplier system in elasmobranch fish is surprising in that the presence of such a system in both birds and mammals mainly facilitates the ability to form an osmotically concentrated urine (*14*). The precise fluid and solute transfers that occur at each segment in the elasmobranch renal tubular system have not yet been established, and we cannot specify, on the basis of anatomical reconstruction alone, how a countercurrent mechanism would work. Ultrastructural studies indicate, however, that the tubular epithelial cells are diverse within the various segments of the tubules (*15*) and have the potential for active transport of solutes. Integration of such information into a realistic model for fluid and solute transfers is an objective for future studies.

ERIC R. LACY
Department of Anatomy, Harvard Medical School, Boston, Massachusetts 02115, and *Laboratory of Electronmicroscopy, Hannover Medical School, 3000 Hannover, Federal Republic of Germany*
ENRICO REALE
Laboratory of Electronmicroscopy, Hannover Medical School
DANIEL S. SCHLUSSELBERG
WADE K. SMITH
DONALD J. WOODWARD
Department of Cell Biology and Physiology, University of Texas Health Science Center, Dallas 75235

References and Notes

1. H. W. Smith, *Am. J. Physiol.* **98**, 296 (1931).
2. ——, *Biol. Rev. Cambridge Phil. Soc.* **11**, 49 (1936).
3. Marine elasmobranchs have mechanisms that allow their body fluids to be nearly isosmotic with seawater, thus preventing major fluxes of water into or out of their tissues. Although both sodium and chloride ions are abundant in the plasma, they are excreted in equally high concentrations (*1*). Urea and, to a lesser extent, trimethylamine oxide account for most of the elevated osmolality of the internal milieu of these fish (*2*). Urea is conserved since about 90 percent of that filtered through the glomerulus is reabsorbed within the nephron (*2, 4*). Plasma concentrations of urea range between 300 and 600 mmol per kilogram of water, depending on the species, and the final urine concentration of urea is only about 50 mmol per kilogram of water (*1, 2, 4*). Although both active transport and passive diffusion have been postulated as underlying this tubular urea absorption (*2, 5*), the process has not been analyzed in detail.
4. R. T. Kempton, *Biol. Bull. (Woods Hole, Mass.)* **104**, 45 (1953); E. K. Marshall, *Physiol. Rev.* **14**, 133 (1934).
5. H. R. von Baeyr and J. W. Boylan, *Bull. Mt. Desert Isl. Biol. Lab.* **13**, 121 (1973); J. W. Boylan, *Comp. Biochem. Physiol.* **42**, 27 (1972); P. Deetjen, D. Antkowiak, J. W. Boylan, *Bull. Mt. Desert Isl. Biol. Lab.* **12**, 28 (1972); H. Stolte, R. G. Galaske, G. M. Eisenbach, C. Lechene, B. Schmidt-Nielsen, J. W. Boylan, *J. Exp. Zool.* **199**, 403 (1977); B. Schmidt-Nielsen, K. Ullrich, G. Rumrich, W. S. Long, *Bull. Mt. Desert Isl. Biol. Lab.* **6**, 35 (1966); B. Schmidt-Nielsen, B. Truniger, L. Rabinowitz, *Comp. Biochem. Physiol. A* **42**, 13 (1972).
6. The elasmobranch nephron is one of the most complex among vertebrates (*4, 5*). Tissue maceration and tubular lumen dye injection studies have provided some details of nephron configuration, but the highly tortuous path of the tubule has made a complete description of the nephron difficult. This in turn has prevented accurate tubular fluid sampling by micropuncture since the puncture site cannot be conclusively identified. Although a peculiar parallel arrangement of some nephron segments was reported in earlier studies (*5, 7*), each proposed model of this part of the elasmobranch nephron has been inferred either from partial injection of nephrons or from time-lapse cinematography, which does not allow determination of the exact location of tubules that may appear to be adjacent at the low magnifications used. All of the models have therefore been significantly different (*2, 5, 7*). Thus, the tubular site for this urea conservation has awaited more detailed morphological analysis.
7. E. Borghese, *Z. Zellforsch. Mikrosk. Anat.* **72**, 88 (1966); P. Deetjen and D. Antkowiak, *Bull. Mt. Desert Isl. Biol. Lab.* **10**, 5 (1970); P. Deetjen and J. Boylan, *ibid.* **8**, 16 (1968); R. T. Kempton, *J. Morphol.* **73**, 147 (1943); *ibid.* **111**, 217 (1953); E. R. Lacy, B. Schmidt-Nielsen, R. G. Galaske, H. Stolte, *Bull Mt. Desert Isl. Biol. Lab.* **5**, 54 (1975); K. Thurau and P. Acquisto, *ibid.* **9**, 60 (1969).
8. Fixed renal tissue was embedded in epoxy resin, and 1-μm-thick sections were stained with toluidine blue and photographed with the light microscope.
9. There was an average of 5.7 μm of tissue between successive photographs, which were enlarged approximately 250 diameters.
10. The outer contour of each tubule profile in each photograph and of the peritubular tissue sheath in every fifth photograph was traced on a digital graphic tablet with a hand-held stylus. The tablet drove a user-defined cursor for digitized images which were stored in a 16-bit minicomputer (Data General Eclipse S/130); this was interfaced to a microprogrammable graphics processor (Adage 3000) for rapid image construction and analysis. Microcoded programs were used for rapid vector generation and smooth-shaped polygon filling of high-resolution images (1000 by 1000 pixels). Video graphics images were used to align contours from serial sections. The graphic data base was used to reconstruct long tubes from individual contours. A set of contours was then subjected to a surfacing routine, which generated triangles between the contours (*11*). By combining output of a lighting algorithm with perspective calculations, polygon fill, and Z-buffer hidden-surface techniques, we constructed a view of the three-dimensional surface of the skate nephron. Photographs of the three-dimensional images were made directly from the video screen. See *Computer Graphics World* (Pennwell, Tulsa, Okla., May 1983) and *Electronics Imaging* (Morgan-Grampion, Boston, 1983), pp. 26–32, for color illustrations of a range of familiar biological objects imaged by these methods.
11. H. Fuchs, Z. M. Kedem, S. Uselton, *Commun. ACM* **20**, 693 (1977).
12. All segments of the nephron in the straight part of the peritubular sheath were 285 μm long since the five tubules were in strictly parallel alignment. Loop I did not extend to the terminus of the peritubular sheath, whereas loop III (Fig. 1a) not only extended to the terminus of the peritubular sheath but was continuously wrapped around loop I and the early distal tubule; this accounts for its length of more than 8 mm. The early distal tubule took a virtually straight course in the convoluted part of the sheath and is thus only slightly longer than the sheath itself.
13. J. R. Clapp, *Am. J. Physiol.* **210**, 1304 (1966); W. E. Lassiter, M. Mylle, C. W. Gottschalk, *ibid.*, p. 965; F. Roch-Ramel, F. Chomety, G. Peters, *ibid.* **215**, 429 (1970).
14. K. Schmidt-Nielsen, *Animal Physiology: Adaptation and Environment* (Cambridge Univ. Press, New York, 1975).
15. The following successive segments beginning at the renal corpuscle can be observed in each nephron: (i) the neck segment consisting of cuboidal cells with numerous cilia; (ii) proximal segments with epithelial cells characterized by a brush border; (iii) an intermediate segment, with six morphologically distinct portions; and (iv) the distal segments (early distal in the peritubular sheath and late distal in the connective tissue outside the sheath). The distal segment merges into the collecting duct, which extends to larger collecting tubules.
16. We thank C. Colette, L. Trakimas, F. Köhler, G. Pugh, and B. McOwen for technical assistance. Supported by the Alexander von Humboldt Foundation and the National Institutes of Health (grant AM-06345) (E.R.L.), Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 146) (E.R.), the Biological Humanics Foundation (D.J.W.), and by the National Institute of Alcohol Abuse and Alcoholism (D.S.S.).

7 September 1984; accepted 1 November 1984

In press, more or less, for G. Adelman (ed.) Encyclopedia of Neuroscience

Brain Size

Harry J. Jerison

The gross weight or volume of an adult mammalian brain is a natural biological "statistic" that estimates fundamental between-species "parameters" of structure and function. Among the relationships that have been discovered, the best established and most important is between brain weight and the surface area of the cerebral cortex. It shows that brain size may be used to estimate the total neural information processing capacity of a species. Other quantitative features of the brain that are related to brain size include the number of cortical neurons, average cortical thickness, average neuron density, extent of axodendritic arborization, neuron/glial ratios, and various neurochemical measures.

These are between-species relationships. Individual differences in brain size within a species are usually unrelated to fundamental structural and functional differences, except during early development or when there is pathology. As Darwin discovered, however, domesticated animals usually have smaller brains than their wild progenitors, and in recent years small and large brained populations of mice have been successfully bred to provide important information on the genetic controls of brain structure and correlated behavior.

Ecological and energetic constraints ~~apparently~~ limit brain size and neural processing capacity. But within those constraints there are important variations among animal species, which ~~and the evolution of these variations~~ have been studied in both living and fossil species.

SUBHEAD Individual (within-species) differences in brain size

Human brains reach about ~~95~~ 90 per cent of adult size, which is about 1300 g (standard deviation about 130 g), by the age of five. At birth the human brain weighs about 350 g. Neonatal cranial circumference between 32 and 36 cm is considered normal; smaller or larger heads are usually associated with mental retardation due to microcephaly or hydrocephaly. For incompletely understood reasons, the brain shrinks about 10 per cent during senescence. Cell death is involved, but the loss of neurons has been confirmed in only a few structures, such as hippocampus, which are too small a fraction of the entire brain to account for the entire weight loss. It is likely that some of the shrinkage is related to dendritic degeneration.

The coefficient of variation of brain size is ~~slightly~~ ~~less~~ slightly less than 10 per cent in mammals. Male human brains average about a standard deviation larger than female human brains, a difference that may be related to body size differences; sex differences in brain size occur in most species in which there is significant sexual dimorphism in body size. Differences among populations of the same species can be produced by breeding programs; realized heritability of brain size in mice in such programs has been about $h^2 = 0.6$.

SUBHEAD Brain size as a natural statistic

Cortical surface area (S) in mammals is related to brain size (E) as shown in Fig. 1. The line is a principal axis fitted to the data for 48 species of

Figure 1 about here

mammals. Its equation (as a power function) is:

$$\underline{S} = 3.74\ \underline{E}^{0.91} \qquad (1)$$

The correlation between the logarithms of cortical surface and brain weight for the data of Fig. 1 is $\underline{r} = 0.995$, or 1.00 to two significant figures.

So high a correlation suggests an almost deterministic relationship. It is, therefore, important to note that some of the small differences among species evident in Fig. 1 are statistically significant. For example, on the average, humans have somewhat less cortical surface than expected, and dolphins (Tursiops truncatus) have more. The value of the exponent in Eq. 1 (the slope of the line in Fig. 1) is also important. That it is 0.91 ($\pm$.02) rather than 2/3 means that there is a change in shape in larger as compared to smaller brains, and the high correlation means that the change is orderly. It is reflected in the appearance of convolutions. The human brain is less convoluted than expected in mammals, and convolutedness is primarily related to brain size rather than to something like intelligence.

About 95 per cent of the variance in the volumes of 15 major brain structures (such as cerebellum and diencephalon) in mammalian species may be accounted for by a gross brain size factor. Brain size may thus provide useful estimations of the volumes of these larger structures by between-species regression analysis, and deviations from the estimations may be used as indexes of specialized developments among species.

Among microscopic and molecular measures of the mammalian brain, the average number of cortical neurons per unit of volume has been reported to be a function of the -1/3 power, and the acetylcholinesterase concentration decreases with the -0.2 power of brain size. The average length of dendritic arborization and the average number of glial cells per unit

volume have been reported to be functions of the +1/3 power of brain size.

There is a growing consensus that the neocortex is remarkably uniform in its structure in that the number of neurons under a given area of cortical surface may be the same throughout the brain in all mammalian species. (Primate visual cortex has about twice as many neurons per unit area as do other cortical areas, but it appears to be the only exception to the generalization. Visual cortex in other mammals investigated thus far is similar to other cortical areas in their brains.) Such uniformity is consistent with current evidence that much of the neocortex is organized into structurally similar functional columns. Taken together with the relation between the thickness of neocortex and brain size, the uniformity implies a neuronal packing density, between-species, that varies with the -.14 rather than the -1/3 power of brain size (data on motor cortex). The margin of error for the lower exponent makes it possibly consistent with data on acetylcholine, but not with older data on neuron density. The older preparations may have been inadequately stained, or differential shrinkage during fixation may explain the inconsistency. The trends in the older and newer studies are the same, of course, although the numbers are not.

SUBHEAD Brain size and processing capacity

Structural uniformity of the neocortex and the surface/ volume data displayed in Fig. 1 are evidence for a close relationship between brain size and information processing capacity. The analysis is a kind of syllogism. If we consider the unit of processing to be a neural columnar module, the brain's capacity to process information should be proportional to the number of modules. Since the number of modules is proportional to the cortical surface area, and since cortical surface area is determined by brain size, processing capacity must be determined by brain size. Because

the analysis is statistical rather than deterministic, it may be better to state that brain size "estimates" rather than "determines" processing capacity.

Processing capacity in the brain can be factored into two components, one related to body size, which is the "allometric" component, and the other a residual that remains when the allometric component is factored out. Statistically, the analysis is like a regression analysis, in which brain and body weights for many species are graphed and analyzed to determine how these measures are related. The regression equation can define an "allometric" factor for a group of species, and the residuals represent encephalization factors for each species. The residual has been used directly as an "encephalization quotient," which can rank species in "extra" processing capacity.

SUBHEAD Allometry, encephalization, and brain evolution

Brain and body weights for the major living classes of vertebrates are indicated in Fig. 2, as minimum convex polygons about the data of each class. The graph may be thought of as a map of evolutionary opportunities in brain/body relations that have been realized in living species. (A few of the many available fossils are shown to illustrate important evolutionary trends.) The polygons are regions in "brain/body space" occupied by living species that are at various grades of encephalization and summarize the present adaptive radiation of vertebrates.

Figure 2 about here

The angular orientation of the polygons, easily imagined as clouds containing data points distributed about regression lines, express the

allometric factor, and these are evidently about the same for each group. The vertical displacements of the higher from the lower sets of polygons indicate the extent to which birds and mammals are more encephalized than the lower vertebrates.

Mappings such as those in Fig. 2 are the framework for interpreting an extensive fossil record of the evolution of the brain and the body. "Fossil brains" are actually endocranial casts for which the cranial cavities of fossil animals were the molds. When such casts are made from the skulls of living species, they look like freshly dissected brains, showing the major gyri and sulci in all living birds and all but the largest brained of living mammals. They are equally revealing in fossils. They are less revealing in lower vertebrates, but are almost always good enough to enable one to estimate total brain size. Hundreds of "fossil brains" have been discovered as natural casts or prepared from fossil skulls, and they provide a record of the actual evolution of the brain. Body size can be estimated from other fossil skeletal remains, and brain/body size data are available for most of the 500 million year history of the vertebrates. Added to the record on living species, these fossils make it ~~is~~ possible to ~~make~~ trace the actual history of the vertebrate brain and its encephalization.

Some of the implications of the data on living species are evident in Fig. 2. For example, it is clear that the basic brain/body allometry is similar in the five vertebrate classes that are represented. It is also clear that at least two major grades of encephalization have evolved which distinguish birds and mammals as "higher" vertebrates and the other three groups as "lower" vertebrates. Interestingly, data on sharks and other cartilaginous fish, omitted from the graph, actually overlap both lower and higher vertebrates. Sharks are relatively large brained compared to other

classes of fish. Jawless fish, such as lampreys, are unusually small brained compared to other fish.

The major conclusions from the fossil evidence are the following:

1. The present lower vertebrate grade of encephalization evolved in the earliest bony fish, amphibians, and reptiles between 350 and 450 million years ago For these vertebrate classes encephalization has remained stable until the present. Since about 2/3 of living vertebrate species are members of these three classes, a "lower vertebrate" grade is the vertebrate norm. 450

2. Variations in encephalization are known within lower vertebrates. The most interesting, perhaps, is that carnivorous dinosaurs were significantly more encephalized than their herbivorous prey. In no sense were dinosaurs "small brained." Their brains were essentially normal for reptiles (Fig. 2).

3. The earliest birds and mammals ~~with~~ had evolved to a higher grade than their reptilian ancestors, representing at least three or four times as much brain as expected for reptiles of comparable body size (points x and a in Fig. 2). The data are from species that lived 150 million years ago, and mammals may have become encephalized 50 million years earlier.

4. Within the mammals there is a good fossil record of the brain, and the history of most orders is consistent with long periods of stability, occasionally "punctuated" by rapid evolution to higher grades. Many familiar species, such as dogs, cats, horses, and hogs, are "average" with respect to encephalization, but there are extremely successful living species (opossum, hedgehog) that are at the same low grade as mammals of 150 million years ago -- with brains about 1/4 the size in average living mammals.

5. Primates have been brainy throughout the known history of their brains, the past 50 to 60 million years, perhaps doing with their brains what many other species did by morphological specializations. However, encephalization in primates apparently followed rather than accompanied their invasion of major new niches.

6. The highest grade of encephalization is shared by humans and bottlenosed dolphins. The human sapient grade was attained about ~~about~~ 200,000 years ago, whereas cetaceans may have reached their highest grade 18 million years ago. As a phenomenon of the past three to five million years, hominid encephalization may be unique among vertebrates in its recency.

Most of the evolutionary enlargement of the brain is explained by the evolution of larger bodies and brain/body allometry. Encephalization has not been a major phenomenon in most vertebrates. On the other hand, its appearance in many different and distantly related groups is evidence of some Darwinian 'fitness' for this complex adaptation.

The behavioral correlate of encephalization may be thought of as behavioral capacity, or "intelligence." It is a general rather than specialized capacity, based on the evolution of different brain systems in different species. The evolution of encephalization, therefore, implies the evolution of a variety of intelligences in animals.

MAJORHEAD FURTHER READING

Reviews and symposia

Armstrong E, Falk D, eds (1982): Primate Brain Evolution: Methods and Concepts. New York and London: Plenum

Blinkov SM, Glezer II (1968): The Human Brain in Figures and Tables. New York: Basic Books Plenum

Edelman GM, Mountcastle VB (1978): The Mindful Brain. Cambridge, Mass.: MIT

Edinger T (1975): Paleoneurology, 1804-1966: An annotated bibliography. Adv. Anat. Embryol. Cell Biol. 49:12-258

Hahn ME, Jensen C, Dudek BC, eds. (1979): Development and Evolution of Brain Size: Behavioral Implications. New York: Academic Press

~~Harvey PH, Clutton-Brock TH, Mace GM (1980): Brain size and ecology in small mammals and primates. Proceedings of the National Academy of Science, U.S.A., 77:4387-4389. GET LATER REF FROM EVOLUTION NOT EDITED!~~

Hopson JA (1979): Paleoneurology. In: Biology of the Reptilia, Vol. 9 Gans C, Northcutt RG, Ulinski P, eds. New York: Academic Press

Jerison HJ (1983): The evolution of the mammalian brain as an information processing system. In: Advances in the Study of Mammalian Behavior, Eisenberg JF, Kleiman DG, eds. Spec. Publ. 7, Amer. Soc. Mammal.

Jerison HJ (1985): Issues in brain evolution. Oxford Surveys Evol. Biol. 2:(in press).

Martin RD (1983): Human Brain Evolution in Ecological Context. James Arthur Lecture. Amer. Mus. Nat. Hist.

Oakley DA, ed. (1985): Brain & Mind. London: Methuen [NOTE: "&" not "and."]

Rockel AJ, Hiorns RW, Powell TPS (1980): The basic uniformity in structure of the neocortex. Brain 103:221-244

Weiskrantz L, ed. (1985): Animal intelligence. Phil. Trans. Roy. Soc. London B308:1-216

Wimer RE, Wimer CC (1983): A geneticist's map of the mouse brain. In Genetics of the Brain. Leiblich I, ed. New York & London: Elsevier

Wimer RE, Wimer CC (1985): Animal behavior genetics: A search for the biological foundations of behavior. Ann. Rev. Psychol. 36:171-218

Monographs

Hofman MA (1984): Towards a General Theory of Encephalization. Amsterdam: Rodopi

Jerison HJ (1973): Evolution of the Brain and Intelligence. New York: Academic Press

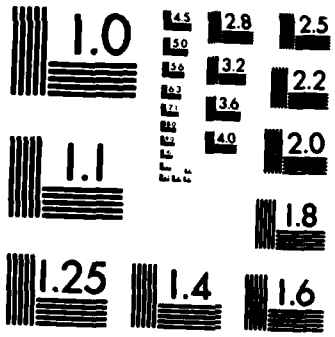
Application of brain size

Ball MJ, MacGregor J, Fyfe IM, Rapoport SI, London ED (1983): Paucity of morphological changes in the brains of ageing beagle dogs: Further evidence that Alzheimer lesions are unique for primate central nervous system. Neurobiol. Ageing 4:127-131

Nellhouse G (1968): Head circumference from birth to eighteen years. Pediatrics 41:106-114

AD-A188 889 PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP HELD IN COLLEGE..(U) TEXAS A AND M UNIV COLLEGE STATION R B LIVINGSTON AUG 85 3/5

UNCLASSIFIED DAAD-17-85-G-5042 F/G 6/5 NL



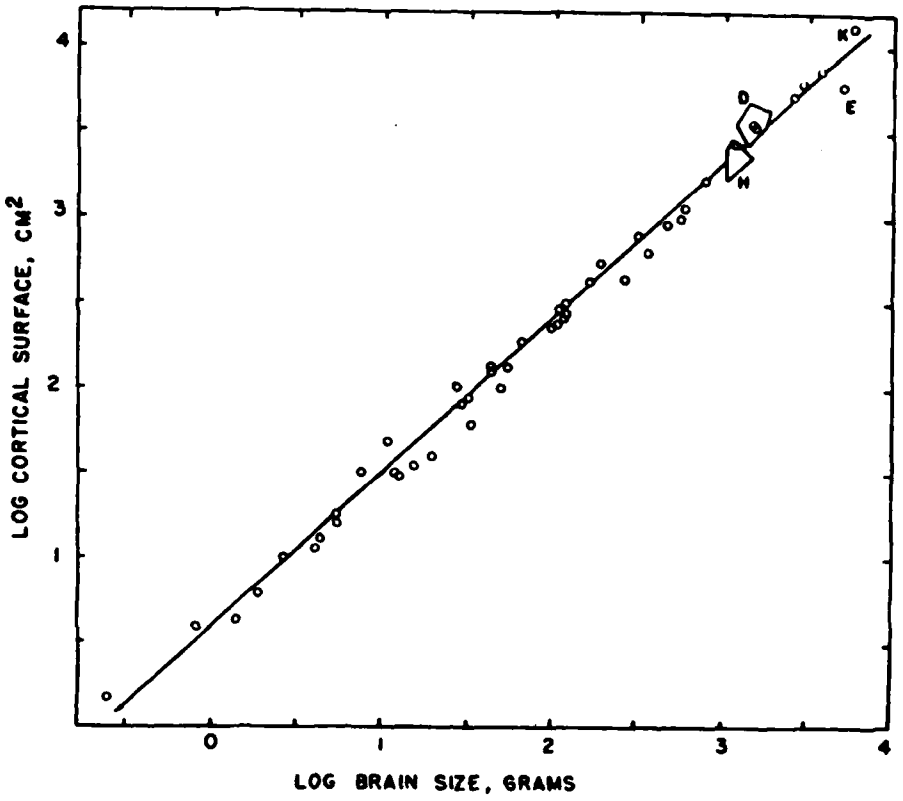
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

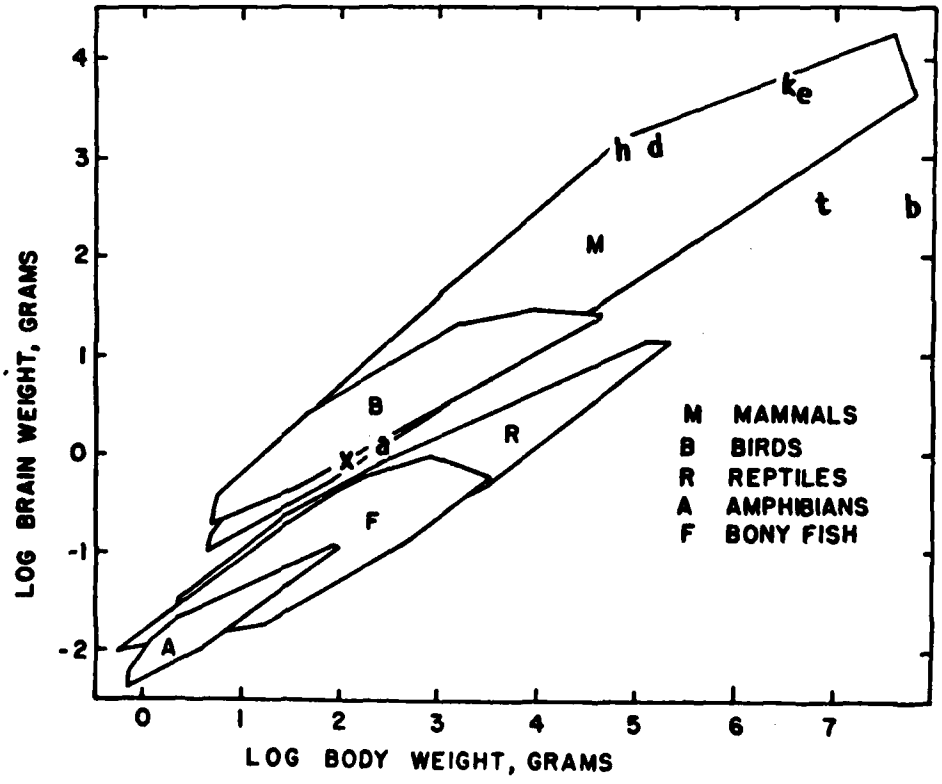
Legends for Illustrations

Figure 1. Cortical surface as a function of brain weight in ~~48~~ 49 species of mammals. Data in small polygons at upper right are on humans (H) and dolphins (D, *Tursiops truncatus*). Other labeled points are a killer whale (K, *Orca*) and an Indian elephant (E). This is a very heterogeneous sample of species, including insectivores, rodents, primates other than humans, ungulates, carnivores, edentates, marsupials, and monotremes.

Figure 2. Brain-body relations in 623 living vertebrate species enclosed in minimum convex polygons for the data of 5 classes. The samples are 309 mammals, 180 birds, 46 bony fish, 40 amphibians, and 48 reptiles. Additional data labeled by lower case letters are the four species labeled in Fig. 1 (d, e, h, and k), 2 dinosaurs (*Tyrannosaurus*, t, and *Brachiosaurus*, b), a 150-million year old mammal (*Triconodon*, x) and the earliest bird, also about 150-million years old (*Archaeopteryx*, a). Note the conservative picture of encephalization suggested by the data: The dinosaurs would extend the living reptilian polygon but would not force either an upward or downward displacement, and the earliest "higher" vertebrates were either at or only slightly below the lower margins of the living mammalian and avian polygons and already above the reptilian polygons.

3.4.12


LOG CORTICAL SURFACE, CM²
LOG BRAIN SIZE, GRAMS
K
D
E
M
0
1
2
3
4


LOG BRAIN WEIGHT, GRAMS
LOG BODY WEIGHT, GRAMS
k
e
h
d
t
b
M
B
R
x
a
F
A
M MAMMALS
B BIRDS
R REPTILES
A AMPHIBIANS
F BONY FISH
-2
-1
0
1
2
3
4
5
6
7

To appear in S. Bacharach, ed., Proc. 9th Internat. Conf. on Information Processing in Medical Imaging. The Hague: Martinus Nijhoff, to appear, 1985.

From Medical Images to the Biometrics of Form

Fred L. Bookstein
Center for Human Growth and Development
University of Michigan
Ann Arbor, Michigan 48109

Within biometrics there is a subfield, morphometrics, for analyzing the geometric forms of organisms. Throughout biology and medicine it is useful to know whether two samples of organs or organisms have the same typical form, and, should they differ, to describe their differences; to indicate the shape changes involved in a cycle such as respiration or heartbeat, or accompanying growth over a lifespan; and to characterize the typical deformation that is a deformity or disease, and its response to therapeutic intervention.

Morphometric studies like these draw information from two sources: biological homology and geometric location (Bookstein, 1982). A biological homology is a spatial or developmental correspondence between individuals, a correspondence among definable structures or "parts"—separate bones, nerves, muscles, and the like. In the context of morphometrics it becomes a homology map, a correspondence not of parts to parts but of points to points. For any choice of point or curve upon or inside any particular form, the homology map associates biologically acceptable counterparts, the homologues of the point or curve, on all the other geometric forms in the data set. Morphometrics studies the empirical geometry of homology—variation in the relative locations of sets of homologous points over a sample of forms.

We generally sample this homology map at landmarks, points whose correspondence from form to form is determined with relative ease. The particular nature of the landmarks used in a study varies with the biometric context: the abutment of two bones, the origin of a valve, a reliable "corner" of sharp curvature, or a metallic implant or marker.

In this essay I shall review recent work (Bookstein, 1984-1986; Bookstein et al., 1985) dealing with the analysis of changes in landmark configurations interpreted as homology maps. This thrust, the construction of shape change as deformation, was first suggested by D'Arcy Thompson in 1917, but its practical application required high-speed computation (Bookstein, 1978). There are two themes to be elaborated. The comparative information we seek can be extracted in a particularly efficient way by the algebraic manipulation of symmetric tensors. And this information may be displayed intuitively, even aesthetically, by computer-generated quantitative diagrams faithful to the data.

Prospectus. I begin by discussing the basic unit of morphometric analysis—the shape change of a triangle of landmarks—and demonstrate its use for describing a normal

human heartbeat. This approach is extended to landmark configurations more complex than triangles by the method of biorthogonal grids, which aids in the interpretation of some canine coronary data. Passing from consideration of single shape changes to the statistical analysis of large groups, I explain how to study populations of shape changes by multivariate statistical procedures. Examples include a simulation of texture change and studies of growth and of deformity in the human head. My concluding remark emphasizes the unexpected simplicity of the biometric information that we ought to be extracting from images.

Principal Strains Computed from Triangles of Landmarks

The basic unit for the study of biological shape change is a homologous pair of triangles of landmarks, Figure 1a. In the absence of other information we may take the homology map sampled by these limited data to be geometrically uniform, as indicated clearly in the transformation grid after the style of D'Arcy Thompson (1961), Figure 1b.

The visual impression this leaves depends on the orientation of the square grid upon the starting form; but this orientation is arbitrary and irrelevant. We draw the transformation more judiciously by its effect upon a collection of lines in all directions, Figure 1c. The deformation we are studying, driven by the displacements of those landmarks at the corners, deforms these segments into others which divide the edges in the same fractions. That is, the deformation takes edges to edges, median lines (dividing the opposite sides in the ratio 50:50) to medians, and so on.

To fully describe a change of form, it is sufficient to know ratios of lengths of corresponding lines in the two triangles. These dimensionless ratios are called strains or extensions. It is easiest to compute them implicitly: they are the lengths into which originally equal lengths—diameters of a circle—are deformed. Let us draw a circle, then (Figure 1d), and the oval into which the uniform shear takes it. Under the assumption of homogeneous (linear) transformation, this oval is an ellipse, precisely. Being an ellipse, the image of the circle has two axes of symmetry, which lie at 90 degrees. One is the largest diameter of the ellipse, one the smallest. The diameters of the circle which transform into them are likewise at 90 degrees.

Recall that the diameters of the ellipse embody the strain ratio as a function of direction. One of the axes of the ellipse is therefore the direction of *greatest* strain, the greatest rate of change of length, and one is the direction of *least* strain. The diameters that were mapped into them are determined by corresponding fractions of intersection along edges of the triangles. In Figure 1e we have drawn these diameters, and in Figure 1f a sketch of their straightforward measurement as transects across the triangles. These axes are called the principal axes of the deformation, and the rates of change of length along them are the principal strains. Together they completely describe the change in form of this triangle of landmarks. The area of the triangle changes by the product of the strains—1.14x0.62 = 0.707—while the most

Figure 1

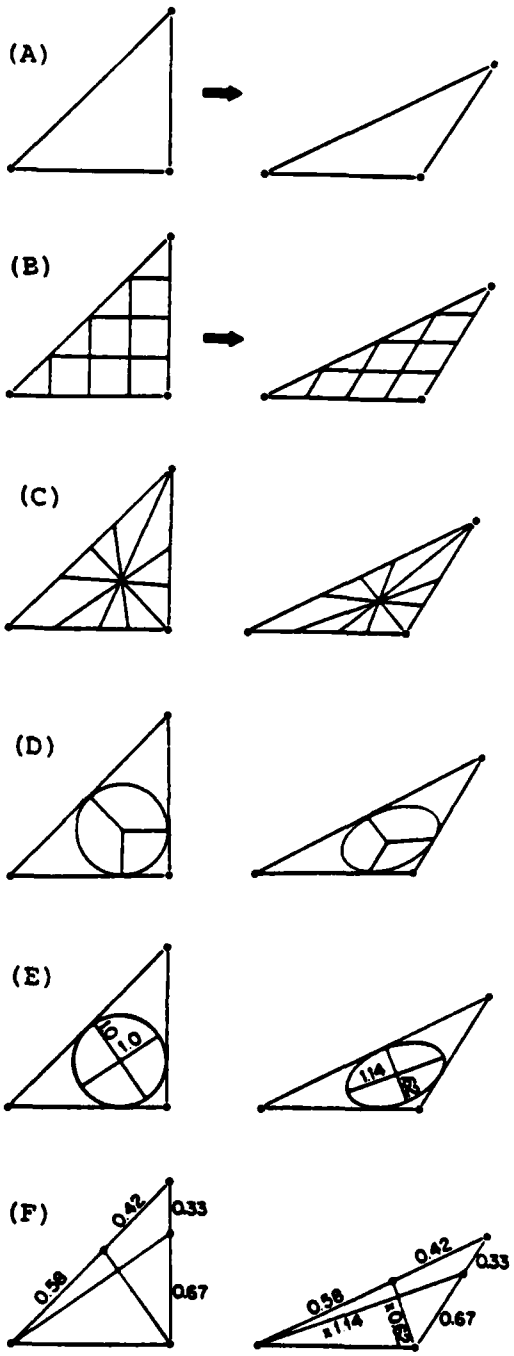


Figure 2

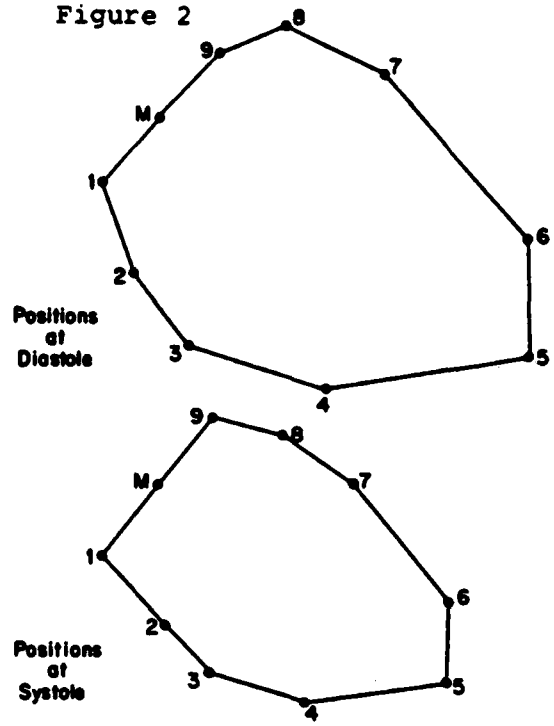
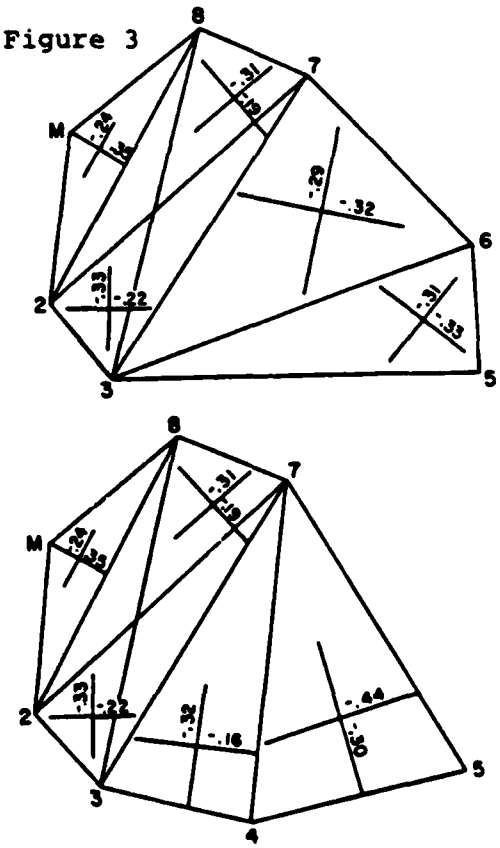


Figure 3



sensitive descriptor of shape change is the proportion between lengths measured in these two directions, which changes by the factor 1.14/0.62 = 1.84. Note that we are not "measuring" the triangles separately at all. The shapes themselves have merely been archived; no preconceptions of specific variables have interfered with the technique's construction of optimal descriptions of change.

The analysis of shape change in this way is not new. The geometrical object just introduced is a representation of homogeneous deformation that is familiar to the mathematician or engineer: a symmetric tensor formally independent of any choice of coordinate system a priori. In this form it appears frequently in mathematical discussions of growth and in cartography, geology, and other sciences of position. In the present application, the "interior" of a morphometric triangle is not a homogeneous biological substance but an arbitrary mosaic of tissues, fluid, or air. The shape change of this abstract interior provides the most useful description of changes in the configuration of its vertices, the real biometric data.

Two Analyses of One Cardiac Cycle

The description of shape change by symmetric tensors provides a very interesting visualization of a single human heartbeat. The data for this example (Figure 2), taken from a published figure (Ingels et al., 1981, Figure 3), locate nine tantalum screws implanted in an otherwise normal human left ventricle during coronary bypass surgery. The view is 30 degrees right anterior oblique; the apex (bottom) of the heart is at marker 5; the base (where the left ventricle empties into the aorta) spans markers 1 and 9.

The method of Figure 1 was used to describe the deformations from diastole to systole of various triangles of these implants. From the computed tensors derives the report of Figure 3, in which the principal strains for the deformation of the triangles shown are drawn within their forms at diastole. The coordinate systems of the separate configurations are irrelevant to the conclusions we draw.

Figure 3 shows that triangles 3-7-6 and 3-6-5, totalling half the area of the ventricle, contract nearly uniformly (by some 31%±2% in every direction). The displacement of marker 4 from marker 7, along the normal to the ventricular contour at marker 4, contracts at the same 30% rate. Relative to the uniform contraction, marker 4 is displaced only tangentially, away from marker 3 and toward marker 5. (Some of this heterogeneity is surely due to twisting of the heart about the projection plane.)

This same contraction of about 31% persists quite far from the apex. In triangles 2-3-7 and 8-3-7, which overlap in Figure 3, the maximum contraction is at this same rate. The minimum contraction in these triangles, 19% or 22%, can be thought of as a weakening of the 32% by a superimposed extension of markers 2 and 8 outward.

Consider a basal triangle joining markers 2 and 8 to the midpoint of the aortic valve, M in the figures. As the apex of the ventricle contracts, the base expands under hydrodynamic

pressure. This top triangle contracts across its base by 24% (the same 30%, perhaps, corrected for the apparent divergence of the translations at 2 and 8 just noted); but its projection along the axis of the heart *increases* by 35% from end-diastole to end-systole. In proportion to the general change of scale by .68, this height has doubled.

These aspects of the description may be abstracted into the nearly symmetric scheme of Figure 4. Superimposed on a uniform contraction of 31% are outward displacements at markers 2, 1, 9, and 8 as shown, together with a lateral adjustment at marker 4. Note the rotation of the axis of the heart relative to the aortic valve ring.

Biorthogonal grids for the same data. The method of triangles computes one tensor per three landmarks, a tensor supposed to apply homogeneously to every point inside the triangle. However, these triangles overlap—they represent a single coherent set of points, a polygon. The method of biorthogonal grids (Bookstein, 1978; Bookstein et al., 1985) is appropriate for such extended configurations.

The method begins (Figure 5a) by computing a smooth deformation—a version of D'Arcy Thompson's Cartesian grids—extending the boundary correspondence inward so as to homologously relate the interiors of our two polygons of implants. The deformation is displayed by its effect upon a mesh of points which is square in the top (diastolic) form; the positions imputed to these points after the "deformation" which is the heartbeat make up the distorted mesh inside the bottom form. These two meshes correspond point for point, as can be seen by comparing their relationships to the implants, points whose homology from diastole to systole we know quite reliably. The position and orientation of that starting square grid, although arbitrary, do not affect the subsequent computations. Like the deformations of triangles, this is an abstract mathematical model of homology. It does not describe what is "really there" but instead expresses the change of boundary form in a convenient diagram.

From the derivative of this map a principal strain tensor can be computed at every point (Figure 5b). These are the infinitesimal directions corresponding to those in Figure 1e as applied to "very small" triangles. Just as for triangles, these directions, perpendicular inside both the diastolic and the systolic polygons, bear the greatest and least local rates of contraction of mathematical myocardium. Curves can be constructed (Figure 5c) which run parallel to one arm or the other of the crosses at every point through which they pass. These curves constitute a grid orthogonal in both forms, before and after deformation: a coordinate system not beholden to features of the forms separately but customized for the particular shape change which is cardiac contraction. Like the mesh points of Figure 5a, the locations at which similarly placed curves of the grids intersect, top and bottom, are computed homologues: they correspond exactly under the map.

The gross deformation which is the heartbeat is described by the lay of these curves upon the forms, by the principal strains (rates of contraction of length) and their gradients along the curves, and by the products and quotients of the pair

Figure 4

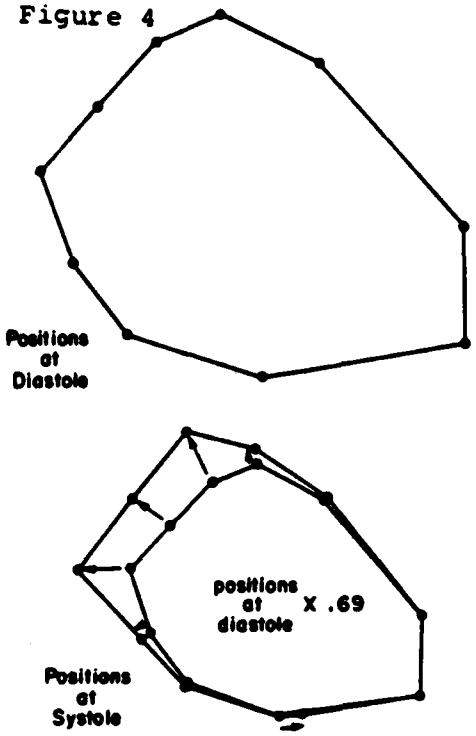


Figure 5(B)

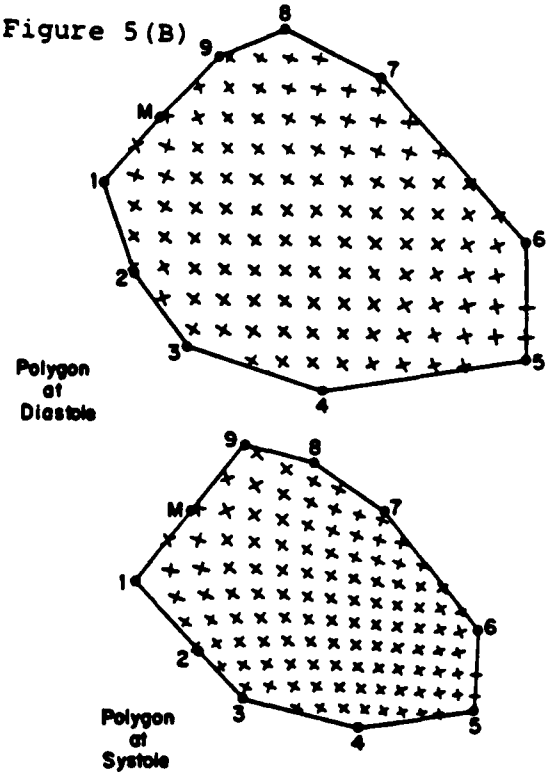


Figure 5(A)

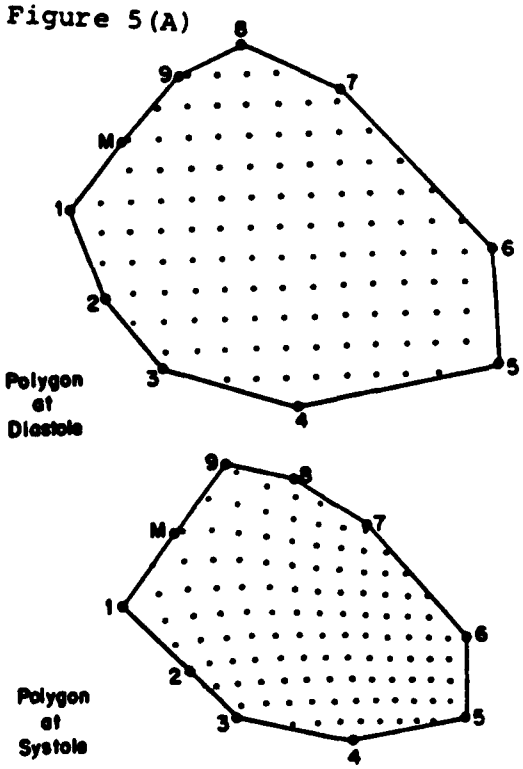
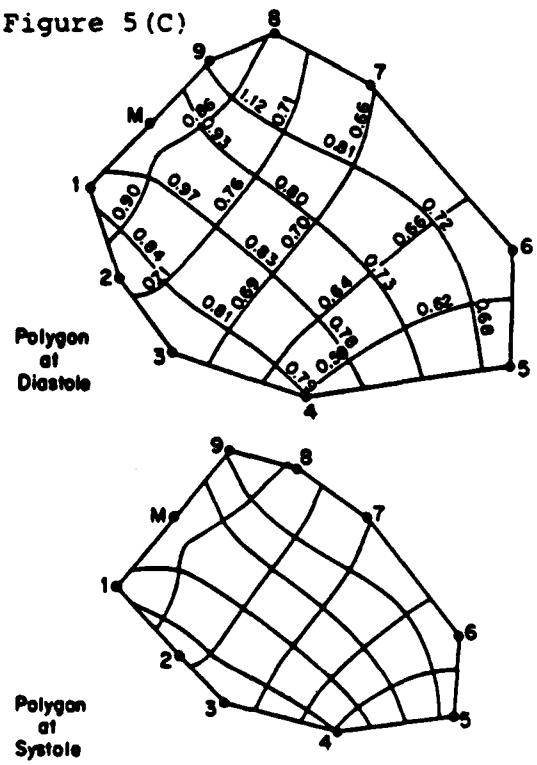


Figure 5(C)



of principal strains at every point (every intersection of curves). In Figure 5c, these strains, indicated within the diagram of diastolic form, are the actual contraction ratios, length in systole divided by length in diastole, for the abstract line-elements upon which they are drawn.

With marker 4 treated as equivalent to all the others, the heartbeat no longer presents the aspect of uniform contraction anywhere. Rather, there is clearly indicated a long axis of least contraction, from M at the base to a point between markers 4 and 5. As we saw in Figure 2, the strain in this direction is graded by a factor of 2, from a compression of some 30% near the apex to expansion near the base. This long axis is one of a system of parallels filling the interior, all showing this same gradient, all slightly curved. Perpendicular to this system are the short-axis curves of greatest contraction, likewise graded from 15%-25% near the base to better than 40% near marker 4, as in Figure 3. Everywhere the little grid rectangles become narrower faster than they become shorter.

The smooth biorthogonal description in Figure 5c is as simple as the discrete analysis of Figure 4. It expresses the same observed change of configuration using a different geometric idiom. For instance, marker 4 now appears to participate homogeneously in a shortening of the septal wall 2-4, a shortening less marked than the long-axis shortening along the free wall from marker 7 to marker 5; this asymmetry is equivalent to the rotation of the valve ring with respect to the heart axis noted in Figure 4.

Localizing Occlusion in Two Experimental Dogs

A principal theme of experimental cardiology is the measurement of coronary occlusion or myocardial infarction from images of the cardiac cycle. In the course of research into the regional analysis of these phenomena, we implanted sets of seven lead shot about the left ventricle (LV) of two experimental dogs. The shot lay in a plane perpendicular to the left anterior oblique (LAO) projection. Each dog was fitted with a balloon occluder of the left circumflex (LCx) coronary artery; dog B bore a second occluder, upon the left anterior descending (LAD) coronary artery. Our biometric polygon is sketched in Figure 6.

Dog A. We imaged dog A in his baseline condition and after sixty seconds of LCx occlusion. For each of the following grids, three consecutive contractions, diastole to systole, were averaged. The contraction at baseline, Figure 7, is represented by a nearly homogeneous grid. In the direction of maximum contraction, the ratio of final to initial length is nearly constant at 0.80 (that is, 20% shortening). Perpendicular to this is the nearly homogeneous direction of least contraction, here at a rate graded from 0.85 (near the apex) to 0.97 near the aortic valve ring.

Figure 9 shows the effect on the heartbeat of occluding the LCx artery. Although the shape of the chamber at diastole has not altered, that at systole has, and so the grids have changed. The occlusion has slightly warped the principal strains of contraction. Far from the region of presumed

Figure 6

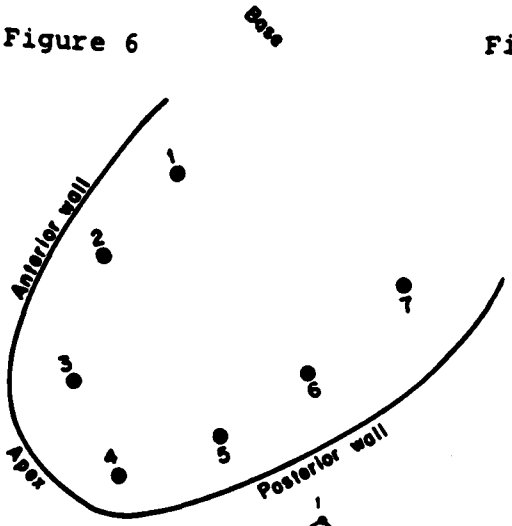


Figure 8

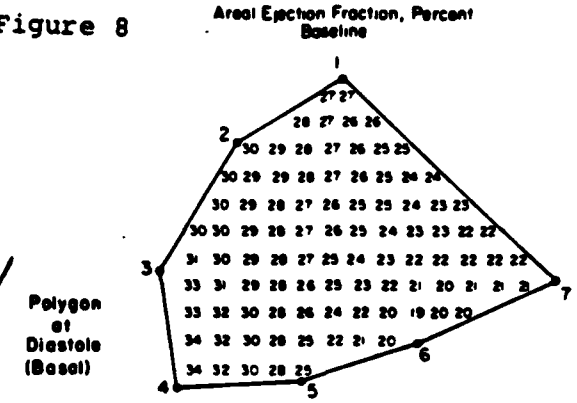


Figure 7

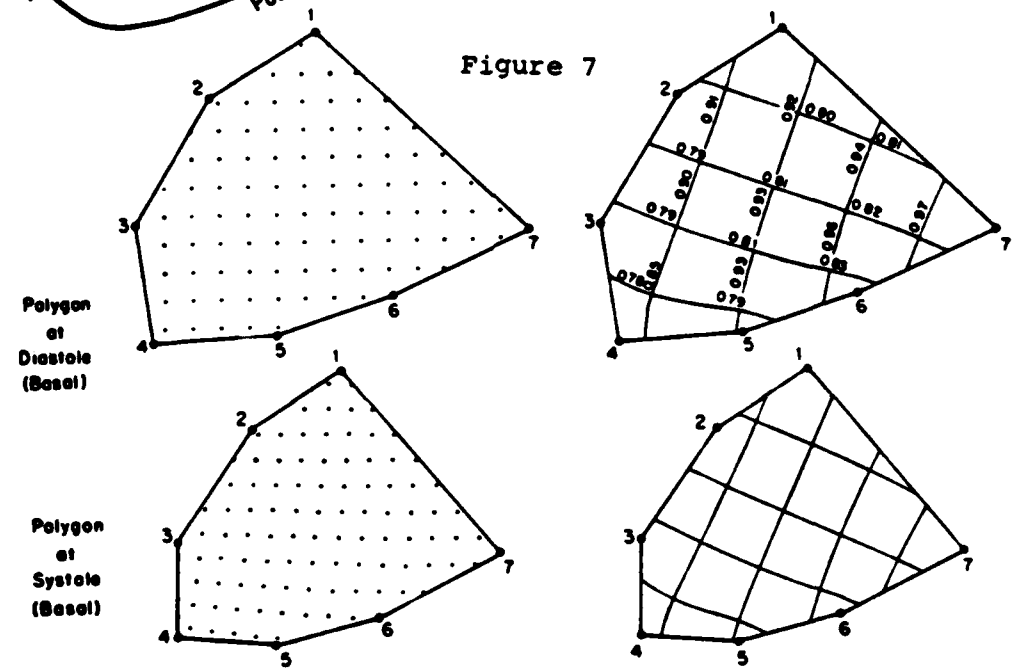
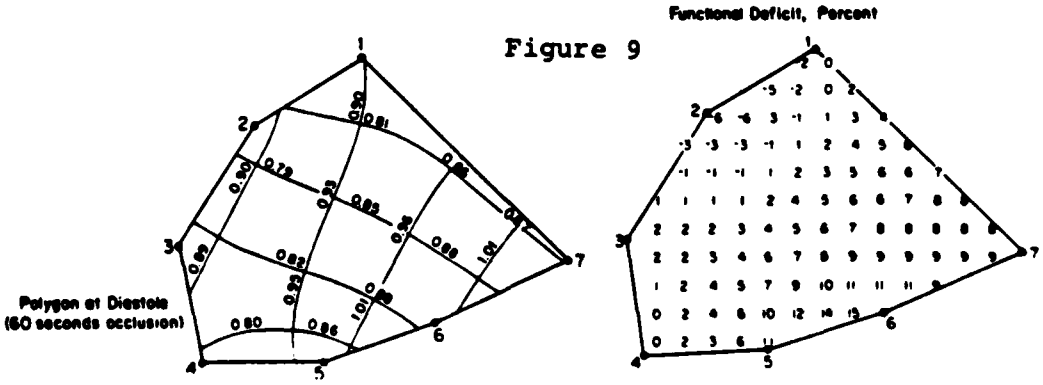


Figure 9



occlusion (arrow), the rates and directions of contraction appear unaltered. Nearby, they seem to be systematically weakened in both principal directions.

Each of these contraction patterns may be represented by its Areal Ejection Fraction (AEF), the difference between 1.0 and the product of the two principal strains at every point. That for the baseline contraction, Figure 8, is quite even except in the vicinity of the apex. The other AEF, for the mean contraction under LCx occlusion, bears a steep gradient from upper left to lower right.

Because the shape of the polygon of implants at diastole has not altered substantially, we are able to map the effect of occlusion as an AEF deficit computed as the numerical difference between the baseline AEF and its value under occlusion, mesh point by mesh point. The map of AEF deficit for this occlusion, Figure 9 right, bears a clear maximum at the center of the myocardium served by the LCx artery.

Dog B. A second dog was affixed with seven shot in comparable positions and with occluders of both the LCx and LAD arteries. The shot were imaged in LAO projection in a baseline condition, during LAD occlusion and recovery, in a second baseline condition, and in LCx occlusion.

This dog's baseline contraction, Figure 10a, while as homogeneous as of dog A, shows a more robust contraction throughout the image. The biorthogonal grids for contraction during occlusion, Figure 10bc, are disorganized relative to that at baseline. Each grid bears a singularity at which contraction is by the same fraction (a mere 7-8%) in all directions. The singularity for the LAD-occluded beat is near the septal wall of the chamber; that for the LCx-occluded beat, in the middle of the free wall. Far from the occluded region, the tensors look like those at baseline.

The Areal Ejection Fractions for the two occluded conditions again demonstrate steeper gradients than the baseline AEF. The AEF deficit plot for the LCx occlusion, Figure 10b, is the same as for dog A, indicating the same focus for the myocardial disturbance. The deficit plot for dog B, Figure 10c, shows the wholly different regional emphasis of the LAD occlusion.

Statistical Analysis

To this point I have spoken of form-change as if it were studied one comparison at a time, visualized as a symmetric tensor, and reported. But one can exploit the tensor formalism much more systematically: it can be made to support all the themes of ordinary biometrics. One can compute averages of shape changes, and investigate their variances and covariances or their dependence upon outside factors. Any of these may be tested for statistical "significance" in the face of the chance variation inevitable in biological studies.

The tie between tensors and biometrics is based upon an aspect of Figure 1e already noted in passing. Recall that the ratio of measured lengths in the principal directions of a shape-change tensor is the shape variable that alters most over the course of a deformation. To pursue the implications of this assertion, let us agree to superpose all the homologous

Figure 10

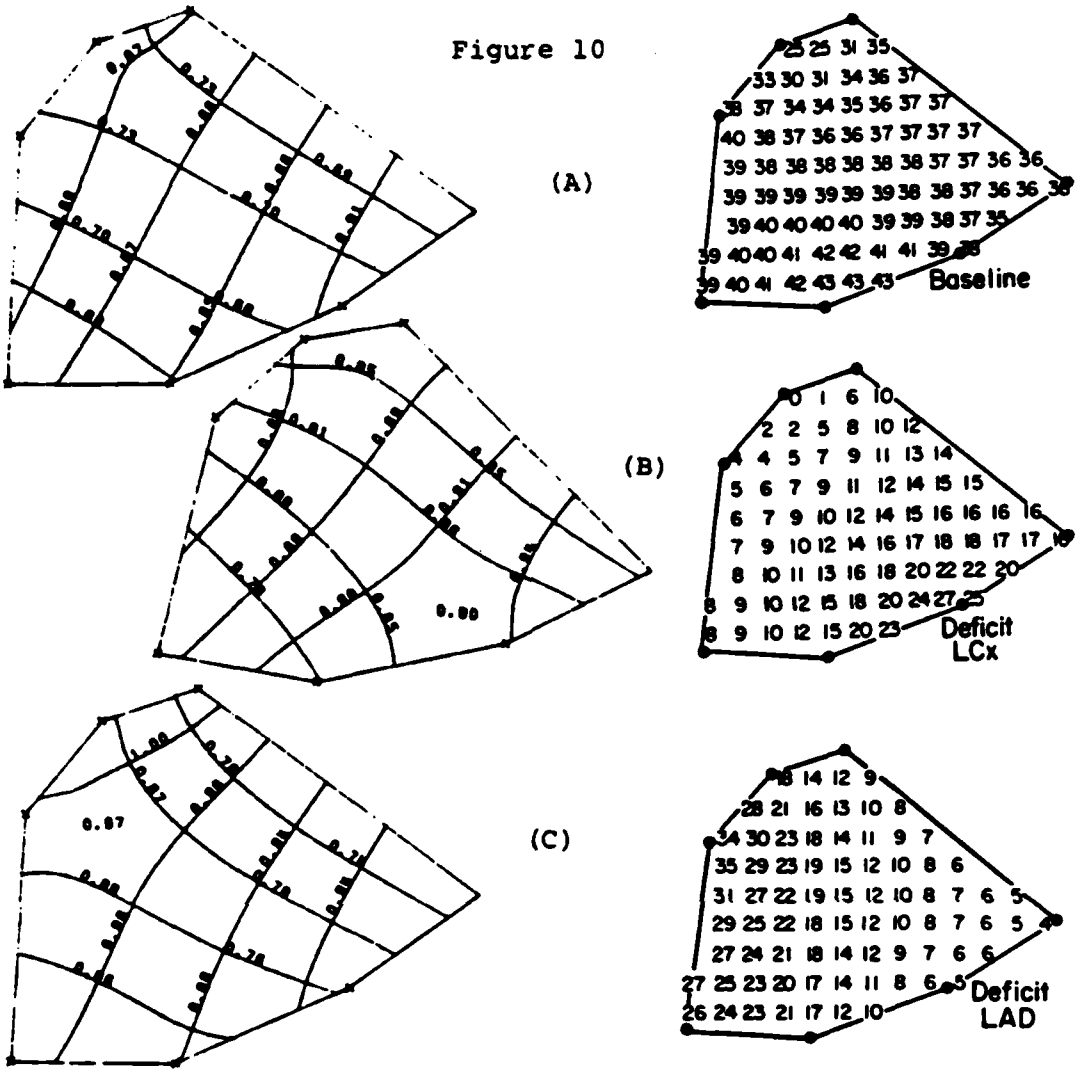


Figure 11

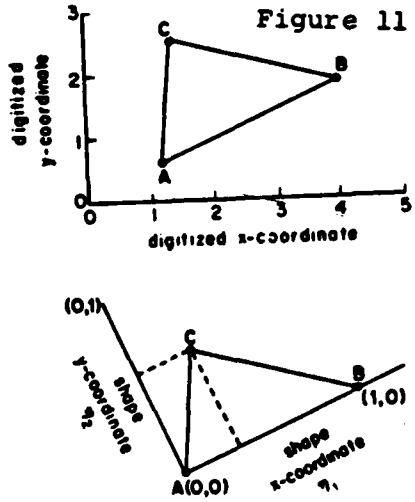
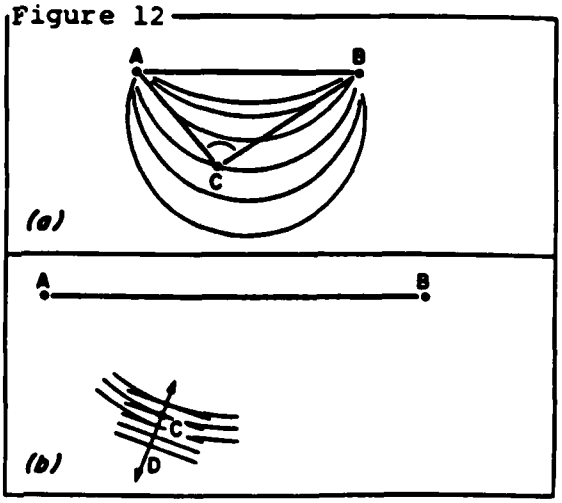


Figure 12



triangles ΔABC of a population in a common position. By change of scale whenever necessary, we shall arrange that landmark A is always put down at the point with Cartesian coordinates (0,0) and landmark B at (1,0), as in Figure 11. In effect we are plotting the triangle ΔABC in a shape coordinate space composed of real and imaginary parts of the complex number $(C-A)/(B-A)$.

Consider any shape variable that can be computed from this triangle of landmarks ΔABC. Figure 12, for instance, illustrates the variable $\angle ACB$. Considering A and B to be fixed in position, any shape variable is constant on some curve through C; the angle $\angle ACB$ happens to be constant along the circle through A, C, and B. Neighboring curves, in this case other circles through A and B, correspond to neighboring, nearly equally spaced values of the shape measure.

In a small region of this plot, this set of curves can be approximated by a family of parallel, equally spaced straight lines, Figure 12b. The shape variable varies fastest perpendicular to these curves, in the direction of the axis D, the *gradient* of the shape variable. The smaller the variation in a population of triangles, the better a shape variable is characterized by the direction of its gradient.

Now consider two points Q and $Q+dQ$ in this space, Figure 13, corresponding to the distinct shapes of two triangles. If every shape variable is a direction in this space, we must be able to find the particular shape variable optimal for describing the difference of these two triangles—the ratio of distances along the two directions of principal strain. This shape variable, which is a *pair of directions* in the original landmark space, is a *vector* in shape coordinate space, the vector connecting the shape coordinate pairs locating the triangles.

One can easily pass back and forth between the representation of shape change by principal axes and this new representation by vectors. The construction (Bookstein, 1984a) is shown in Figure 13. Draw the circle H to pass through Q and $Q+dQ$ with center on the real axis. Let the points at which H intersects the x-axis be denoted $(W,0)$ and $(X,0)$. The angles $\angle WQX$ and $\angle W(Q+dQ)X$ are each inscribed in a semicircle; hence both are right angles. Because the X's are on the real axis, the linear transformation represented here, which leaves (0,0) and (1,0) fixed, also leaves the X's fixed. Then under this transformation the lines WQ and XQ correspond to the lines $W(Q+dQ)$, $X(Q+dQ)$. Because these directions (i) correspond under the transformation, and (ii) are perpendicular in both forms, they must be the principal directions of the transformation to which the construction refers—axes of the ellipse, directions of greatest and least ratio of change of size.

For small changes of shape, the circle through Q and $Q+dQ$ may be approximated by the circle through Q with tangent along the direction dQ there. Then the directions we seek through Q are at ±45 degrees to the bisectors of the angle between dQ and the real axis, Figure 14. Furthermore, if s and t are the strain ratios in the two principal directions, we have, to first-order terms, $s-t = |dQ|/\mathrm{Im}\, Q$. This is the anisotropy of

the transformation, the greatest divergence of specific rates of change (difference of loadings) in any pair of distances.

In this way, tensor biometrics can be applied to whole populations of triangles. The statistical analysis of triangular shapes becomes the statistical analysis of scatters of points in the plane, for which ordinary multivariate maneuvers are quite adequate. It is easy to prove that this algebraic machinery is practically independent of the choice of baseline; it is also straightforward to restore to this analysis the information about size that was divided out when we reduced the triangles to their shape coordinate pairs. See Bookstein (1984b, 1986).

Example 1: Simulation of Change in Texture

As a first demonstration of the statistical analysis of deformations, consider the texture in Figure 15—a scatter of ellipses lacking all landmark information. The ellipses are randomly oriented with axis-lengths varying randomly and independently about a ratio of 2:1. The display in the figure was simulated, but might as easily have represented second-order moments of detected cells or other inclusions in an extended scene.

Each ellipse may be characterized by its shape—the ratio of lengths of its axes—and its orientation. Instead of thinking of them as shapes, however, let us treat each one as the deformation of a circle, so that it may be represented by a vector: its effect on one vertex of an arbitrary but fixed triangle. In effect we are inverting the construction of Figure 1: given the axes and the ratio of strains, to find the vector of displacement of a third point when size is adjusted so as to hold fixed the two other vertices of the triangle. By this tactic, an ellipse of anisotropy δ and principal axis at angle θ to the real axis is plotted as a vector having polar coordinates $(\delta, 2\theta)$. The scene of ellipses is thereby translated into the scatter of vectors shown in Figure 16. The design of the original simulation is plain here: the ellipses lie near a circle of radius corresponding to the expected anisotropy and independent of orientation.

Figure 17 is a modification of Figure 15 by a uniform stretch of 25% in the horizontal direction. The transformation has modestly changed both the shape and the orientation of every ellipse in the scene. Although the result does not appear notably different in directionality, the deformation involved in its construction can be unambiguously detected in Figure 18, the scatter analogous to Figure 16. In this new scatter, the circle is plainly off-center: the average vector of deformation representing the ellipses has been displaced from (0,0). This means that there is a *correlation* between orientation and shape in this new population of ellipses. Ellipses which were aligned up-and-down are now fatter; those which were aligned left-and-right are now longer, hence skinnier. The correlation thus explicitly embodies the additional deformation we have applied to each individual ellipse of the scene.

Figure 13

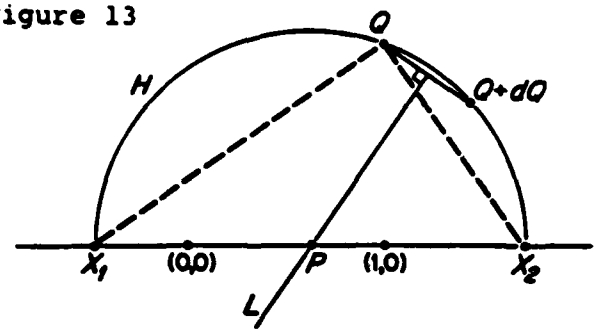


Figure 14

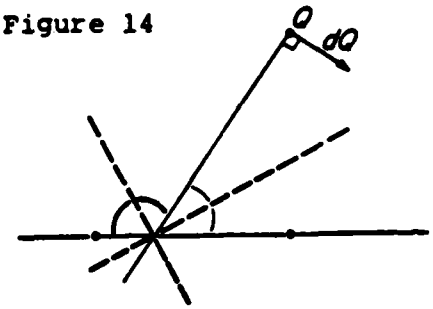


Figure 15

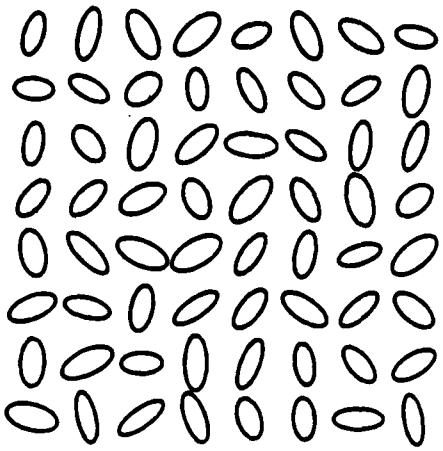


Figure 16

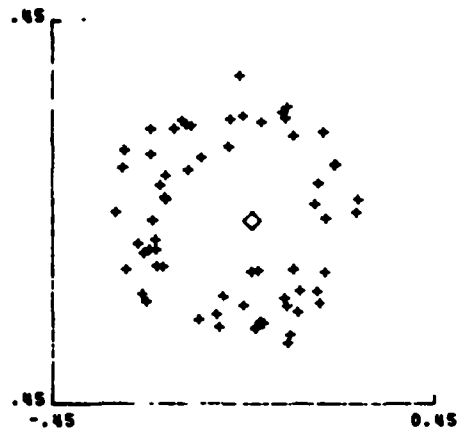


Figure 17

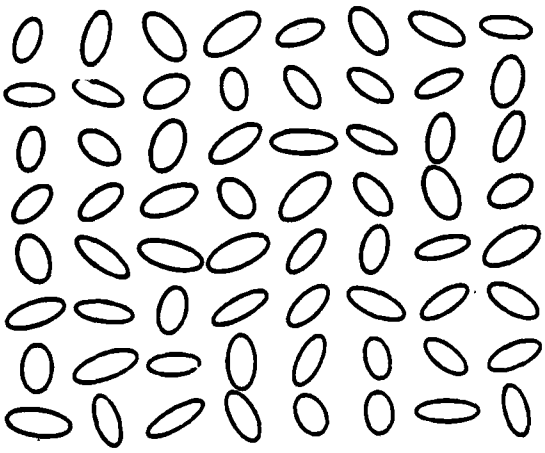
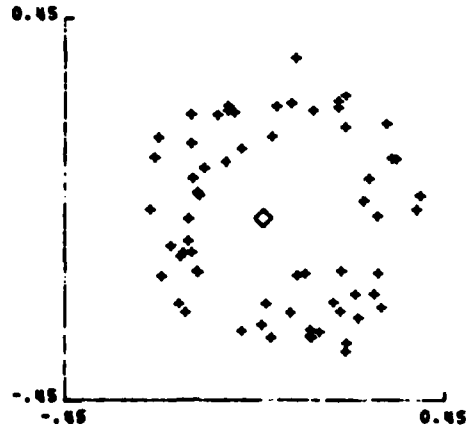


Figure 18



Example 2: Craniofacial Growth, Craniofacial Deformity

Naturally our interest is concentrated in deformations of real landmarks rather than simulated ellipses. In this same manner we may analyze populations of deformations of landmarks by conventional statistical methods applied to constructed points in the shape coordinate plane. The examples following are extracted from previously published analyses of radiological images of human craniofacial form (Bookstein, 1984b).

Cephalometric data. Most craniofacial biometrics begins with x-ray images of bony crania and jaws positioned for exposure in a standard fashion. The subject's head is placed some six feet from the x-ray tube and a few inches from a film cassette; the central beam of x-rays, perpendicular to the film, passes through his ear holes. In the images which result, edges of anatomical structures can be reliably traced in a conventional abstraction of normal anatomy, schematized in Figure 19.

The data for these examples involve landmark locations from cephalograms taken annually in the course of the University of Michigan University School Study (Riolo et al., 1974). The full sample is of about one hundred Ann Arbor schoolchildren followed over various age ranges in the 1950's and 1960's; for a subset of 36 males, there are serial records of the four landmark locations at ages 8 years ± 6 months and 14 years ± 6 months.

The landmarks to which I shall be referring are **Sella**, the seat of the pituitary gland in the middle of the base of the brain; **Nasion**, the deepest point in the curvature of the profile at the bridge of the nose; **Anterior Nasal Spine (ANS)**, tip of the bony palate, just under the nose; **Menton**, point of the chin; **Gonion**, the lateral corner of the jaw, often visible in photographs of jut-jawed males; **Sphenoethmoidal Registration Point (SE)**, the intersection of two shadows (the greater wing of the sphenoid bone and the anterior cranial base); and **Basion**, the frontmost point on the foramen magnum where the spinal column enters the skull. The first six of these landmarks straddle the *splanchnocranium*, that part of the head which deals with breathing, smelling, and chewing rather than with protection of the brain. The locations of these points are manually digizited from pencil tracings of the original x-rays; I know of no means for locating cephalometric landmarks automatically.

Normal growth. The morphometric study of normal craniofacial growth can begin with the large triangle joining the landmarks Basion, Nasion, Menton for the 36 normal Ann Arbor males observed at ages 8 and 14. We fix Basion at (0,0) and Nasion at (1,0), representing the shape of this triangle in these boys by the shape coordinates Q of Menton. The shape change of this triangle from age 8 to age 14 is read in displacements dQ of these coordinates. There results the scatter of vectors dQ in Figure 20. (The "pinheads" locate the earlier of the two positions.) The heavy black vector, connecting the centroid of the earlier shapes to the centroid of the later shapes, represents the mean shape change over this six-year period.

This mean shift of shape coordinates corresponds by the construction of Figure 13 to a pair of distances, along the directions of principal strain, that might have been measured directly upon the x-rays as indicated in Figure 21. The baseline of this triangle, along the cranial base, grows least—by a mean fraction of 9.1% over this six-year period. Perpendicular to it is the direction of greatest average growth, by 17.5% over the same period. This difference of rates—evidence of a mean shape change—is hugely statistical significant. It is, furthermore, substantially correlated with net size change—the more one of these boys grew, the more his chin tended to grow "vertically," away from the cranial base.

There is some statistical regularity to the vectors of change in Figure 20. Chins beginning toward the left or the right (that is, relatively back or forward of the average position) tend to stay left or right; chins beginning relatively high or low (short faces or long faces) do not appear so predictably stable in position. Analogous to the tensor of shape change we have just been discussing, there is a tensor for directional dependence of shape stability, with principal axes of its own. Variation along the horizontal shape coordinate of Q is most predictable, with an autocorrelation of 0.89 from age 8 to age 14; variation along the vertical shape coordinate is least predictable, with an autocorrelation of 0.62 only. That different shape variables may be forecast with different degrees of accuracy is important in studies assessing the effects of orthodontic therapy. That it is so-called *vertical* change which is least predictable is a consequence of the dependence of vertical change upon size change noted in the preceding paragraph. Because the size change of an adolescent boy is essentially unpredictable, those aspects of shape most dependent upon size change will naturally be themselves the most variable. This analysis is extended to additional landmarks in Bookstein (1986).

Characterizing Apert's syndrome. The same design we applied to the longitudinal data of the heartbeat can be used for any other matched design. In particular, we may construe craniofacial deformity as deformation. Any instance of a syndrome may be measured not as a form but as a deformation of the "normal," specifically, of the age- and sex-matched University School Study normative mean. For example, consider Apert's syndrome, exemplified in Figure 22. It is one of the craniofacial *synostoses*, which generally manifest premature closure of the intracranial bony sutures about the maxilla and frontal bone. Apert's syndrome, or acrocephalosyndactyly, shows deformities of the extremities as well. Facially, the syndrome typically include a high, bulging forehead and a short maxilla (upper jaw positioned much higher and further back than normal. Landmark configurations for patients afflicted by this syndrome were lent to me by Dr. Joseph McCarthy from the data bases he maintains at the Institute for Reconstructive Plastic Surgery, New York University.

Six landmarks bound the facial region of interest in the study of Apert's syndrome. Assemble them into a polygon: Sella-SER-Nasion-ANS-Menton-Gonion. We averaged coordinates of

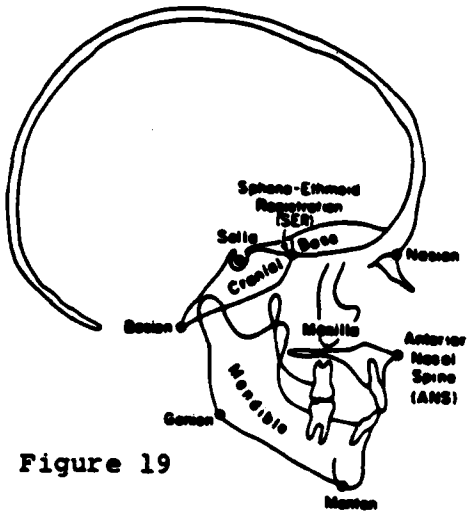


Figure 19

Figure 20

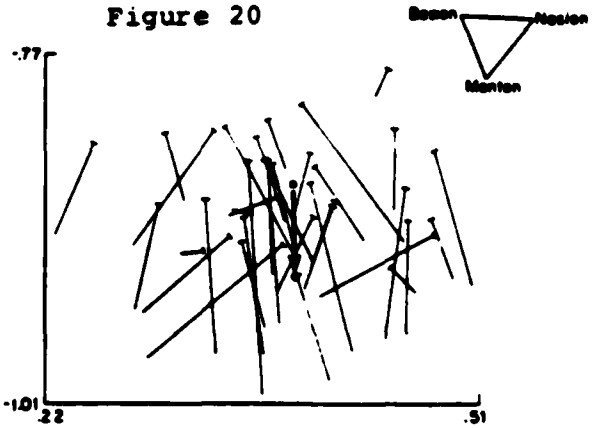


Figure 21

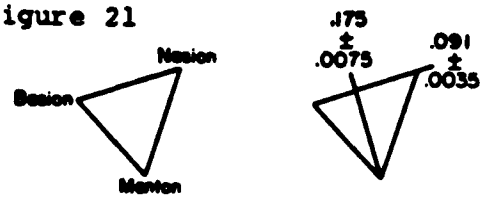


Figure 22

Apert's Syndrome

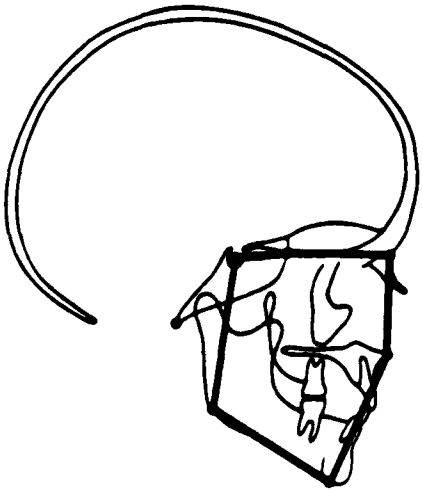


Figure 23

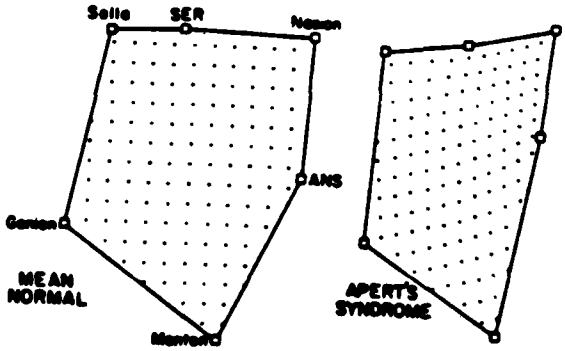
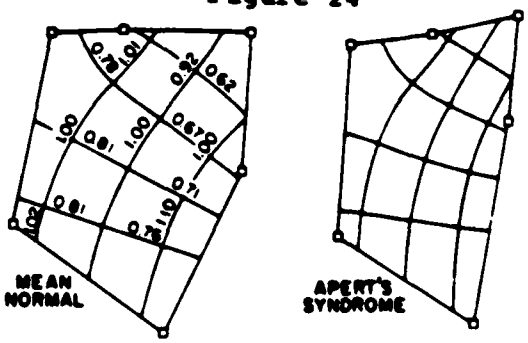


Figure 24



the landmarks in the cephalograms of the patients and of the matched Ann Arbor normal children.

Figures 23 and 24 display the mean deformity of these 11 Apert's cases as the mean *deformation* into the typical case from the normal mean form. Figure 23 uses the diagrammatic style of the interpolated Cartesian grids, while Figure 24 summarizes the same homology map by its biorthogonal grid pair. Some dilatations are indicated, too. In this context, they are ratios of *abnormality of length* segment by segment: deformed length divided by homologous normal length.

The grids show an anteroposterior compression in the typical Apert's form that dominates relative normality of vertical dimension. Lower face height is slightly *larger* than normal, in consequence of the open-bite induced by the syndromes. There results, throughout the front of the face, a disproportion of some 30% with respect to normal shape. The grids are slightly tilted in the vicinity of Nasion, where one linear dimension appears to be most abnormal: the distance from ANS to SE. A more detailed analysis (Grayson et al., 1985) demonstrates that the point SE is nearly the seat of the syndrome; it is displaced within the shadow of the cranial base by more than half its normal distance to the line Basion-Nasion, perhaps in response to hydrostatic pressure from the developing brain.

An assortment of such separately optimal proportions can be submitted to any protocol for multiple discriminant analysis. They are particularly suited to the path-analytic discriminant model of Bookstein et al., 1985, Section 4.3.

Concluding Remark

The medical study of human anatomy, whether in the atlas or in the clinic, used to pursue two purposes jointly: the depiction of the "normal," its mean and its variants, and the authoritative specification of anomaly. Each of these is explicitly a comparative theme. In the modern equation of medical image analysis with computerized image processing, unfortunately the theme of comparison has evaporated, replaced by display of a purely *geometric* scene. The *biological* component of these images is suppressed, especially the information needed for morphometrics: the locations of named (homologous) parts or points in a series of forms.

Without this information, *we cannot compare forms intelligently*. We cannot make up for its omission by any manner of enhanced display. Rather, once a gray-scale image is processed for the extraction of modest biometrical information—the locations of a few carefully selected landmarks—the pixels or voxels are best entirely scrapped, replaced by a pair of abstract polygons, or polyhedrons, and their geometric and statistical derivatives. Tensor morphometric analysis of geometric data supports all the great themes of clinical anatomy—normal means, normal variation, characterization of anomalies—whereas the gray images, however much more realistic, wholly fail to do so.

In my view, there are two or three orders of magnitude too much information in medical images as they are currently displayed. The main need in medical image analysis today is

not for image *processing* at all. Rather, we should concentrate our professional energies upon algorithms for the detection of landmarks (aids to human search, or prescans for human corroboration) and for the flagging of abnormal parts of an extended scene. For studies of form, the remaining gray-scale information is simply to be discarded; for studies of function, it can be analyzed only in the coordinate system supplied by the biometrics of form, the tensor morphometrics proposed in this essay.

Acknowledgement. Preparation of this essay was supported by N.I.H. grants DE-05410 to F. L. Bookstein and DE-03610 to R. E. Moyers. The canine experiment of which one fragment is described here was executed by A. J. Buda and Kim Gallagher under grant HL-29716 to A. J. Buda. The data regarding patients with Apert's syndrome were gathered under grant DE-03568 to J. G. McCarthy.

Literature Cited

Bookstein, Fred L. *The Measurement of Biological Shape and Shape Change*. Lecture Notes in Biomathematics, v. 24. Springer-Verlag, 1978. 191 pp.

Bookstein, Fred L. Foundations of morphometrics. *Annual Reviews of Ecology and Systematics* 13:451-470, 1982

Bookstein, Fred L. A statistical method for biological shape comparisons. *J. Theoretical Biology* 107:475-520, 1984a

Bookstein, Fred L. Tensor biometrics for changes in cranial shape. *Annals of Human Biology* 11:413-437, 1984b

Bookstein, Fred L. A geometric foundation for the study of left ventricular motion: some tensor considerations. In *Digital Cardiac Imaging*, eds. A. J. Buda and E. J. Delp. The Hague: Martinus Nijhoff, 1985, pp. 65-83

Bookstein, Fred L. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, accepted for publication, 1986

Bookstein, Fred L., B. Chernoff, R. Elder, J. Humphries, G. Smith, and R. Strauss. *Morphometrics in Evolutionary Biology. The Geometry of Size and Shape Change, with Examples from Fishes*. Academy of Natural Sciences of Philadelphia, to appear, 1985

Grayson, B., N. Weintraub, F. L. Bookstein, and J. McCarthy. A comparative cephalometric study of the cranial base in craniofacial syndromes. *Cleft Palate J.* 22:75-87, 1985

Ingels, Neil B., Jr., G. T. Daughters II, E. B. Stinson, and E. L. Alderman. Left ventricular midwall dynamics in the right anterior oblique projection in intact unanesthetized man. *Journal of Biomechanics* 14:221-233, 1981

Riolo, M. L., R. E. Moyers, J. S. McNamara, and W. S. Hunter. *An Atlas of Craniofacial Growth*. Monograph No. 2, Craniofacial Growth Series, Center for Human Growth and Development, University of Michigan, 1974

Transformations of Quadrilaterals, Tensor Fields, and Morphogenesis

Fred L. Bookstein

Center for Human Growth
The University of Michigan
Ann Arbor, Michigan

February, 1984

To appear in
Mathematical Essays on Growth and the Emergence of Form
P. L. Antonelli, ed.
University of Alberta Press, ~~1984~~ 1985

I. Introduction

Morphogenetic explanations often invoke geometrical metaphors. The spatial reference may be subtle, as in the classic notion of induction, or explicit, as with polarity or the recent construct of positional information. There is a growing literature, for instance, which argues at length the relation between experimental evidence and a model of polar coordinates for encoding position around a limb. In the course of regeneration, it is argued, the animal intercalates missing structures in the sequence which, modeled as a change of polar coordinate, is shortest. Other uses of geometry have a more conventional topology, as in the attempts [elsewhere in this volume, Editor? If so, please specify] to derive biological compartment boundaries from the resonance geometry of diffusion-reaction partial differential equations.

In the study of shape change, this sort of speculation, whatever its apparent fit to some experimental facts, is in my view an oversimplification of the logical and parametric structure of geometric objects. The "coordinate" of a point in the plane, for instance (Bookstein, 1981), ought to be seen not as a pair of quantities but as a pair of curves through the point specifying sets of other points with which it is, literally, *co-ordinated*, sharing a coordinate value. The manner in which these curves are interconnected in the large is crucial to their effectiveness in explaining biological order. For instance, coordinates which appear radial with respect to one polar center may serve at the same time as azimuths about another center at

some distance, as in Figure 1a; or neighboring isopleths of one coordinate axis, Figure 1b, may prove to be opposite ends of the same coordinate curve. (This pattern plays a major role in root meristem growth in plants; see Schüepp, 1966.) The vector fields of morphogenesis, presumed gradients of an underlying "morphogen" (scalar field or positional coordinate), may have singularities of a structure rich in two dimensions, even richer in three. This has been amply shown by Winfree (1980).

But scalars and vectors are only the zeroth and first levels of a hierarchy of complexity for the *tensor fields* which represent the interactions of multiple geometric parameters at a point. This essay explores implications for morphogenesis of the next level in this hierarchy, the *second-order symmetric tensor* representing two positive strains or other rates in two perpendicular directions. Just as scalars are indicated graphically by decimal numbers or contour lines, and vector fields by little line-elements, a symmetric tensor field is diagrammed as a little perpendicular cross at every point. Coordinate systems based on these tensors have two generic types of singularity quite different from those arising in the study of vector fields. Part IV of this essay will derive the two typical forms of these singularities, and Part V will speculate on their role in morphogenetic explanation. These sections are preceded by some algebraic preliminaries which, beginning with the interplay of tensor and vector descriptions for deformations of triangles, ultimately call our attention to certain transformations of quadrilaterals that demonstrate ~~all~~ the tensor

singularities in which we are interested.

II. The symmetric tensor field as a coordinate grid for deformation

In the biophysical study of embryology several mechanisms are considered that might be described by symmetric tensors: changes in cell-cell contacts, relations between layers of cells, programs for orientation of cell division. When a tensor theory of morphogenesis finally confronts data, any of these might serve as actual encoded morphogen over a region. In this essay I will rely on the simplest embodiment of a symmetric tensor field, its role in describing deformation. Point by point, biological form-change may be described by two rates of linear extension in two directions at 90°.

Such a representation begins with the *homology map*, the "Cartesian transformation" formalized by D'Arcy Thompson (1961). A biological homology is a spatial or ontogenetic correspondence among definable structures or "parts" -- separate bones, nerves, muscles, etc. In the context of mathematical morphology it becomes a homology <u>map</u>, a smooth geometrical *deformation* not of parts to parts but of points to points. For any choice of point or curve upon or inside any particular form, the homology map associates well-defined and biologically acceptable counterparts, the *homologues* of the point or curve, on all the other geometric forms in the data set.

A homology map may be drawn, after the fashion of Thompson, as the distortion of square graph paper into a more general configuration. The map relating a typical pair of four-cornered

regions might look as in Figure 2a. Of course, because data are supplied only at the corners, what is displayed must represent an interpolation formula, realistic, perhaps, but arbitrary. That shown here is the bilinear map described in Part IV; many others are possible (Bookstein, 1978).

At almost every point interior to either quadrilateral, Figure 2b, there is one cross of directions that is orthogonal in both forms according to the interpolated homology map. In the neighborhood of this cross, the transformation consists in independent expansion or contraction of each arm of the cross by separate ratios without change of angle between the arms. In one of these directions, rate of change of length is greatest, and in the other, least, of all directions across the triangle. These directions, the *principal axes* of the deformation, are at 90° in both forms. The rates of change of length along them, computed as dimensionless ratios, are called the *principal strains* or *principal extensions* at the point. The larger will be denoted d_1, the *major* principal strain; the smaller, d_2, the *minor*. The axes and principal strains together make up the *principal cross*, a visualization of the *symmetric strain tensor* (see) familiar from continuum mechanics and engineering.

Viewed in this way, locally the deformation has a *size component* quantified by $d_1 + d_2$: the rate of change of area, unrelated to direction. Complementary to this is a *shape component*, or *anisotropy*, measured as $d_1 - d_2$, the difference of the principal strains. This quantity expresses change over the deformation in the ratio between measured distances along the two

principal axes.

During the course of any change, there appear translations and rotations between larger parts of the form. The deformation model attributes these to the net influence of the local strains, their magnitudes and directions, summated across the form. The integration of these principal strains into extended curves, Figure 2c, assembles the depiction of the tensor field into a recognizable pattern: a pair of curvilinear coordinate systems, each bearing a constant angle of 90° in both forms. These make up the *biorthogonal grid pair* representing the mapping. Upper and lower grids correspond, intersection for intersection, according to the interpolating map we are using in Figure 2a. The selection of these curves is arbitrary, but their orientation is not: each curve is precisely parallel to one arm or the other of the principal cross at every point through which it pass. These curves are at 90° in both forms wherever they intersect: they constitute a coordinate system customized for this particular shape transformation. A relative rate of extension may be read from the grids as the ratio of the two homologous lengths (left and right) cut off along one curve by successive transects with the perpendicular system. The reader interested in further study of these coordinate systems and their application to empirically observed changes should consult Bookstein, 1978, or Bookstein et al., in prep.

This essay takes the grids to represent real morphogens rather than, as Thompson averred, mere expression of a system of "forces" at a distance. In other words, the cell, organism, or

tissue is assumed to generate its change of form actively, not passively, within the region under study. (For a related model, see Jacobson and Gordon, 1976.)

Shape change of a single triangle. For the parametric treatment of deformation grids, we require some algebraic maneuvers relating to the apparent relative displacements of corners of the form over the course of a deformation. Explanations are simplest if we begin with a single cross representing a strain homogeneous in its little region. We may refer this to a set of three vertices, Figure 3a, because a change in shape of any configuration of three homologous points can be modeled as a homogeneous deformation of the interior of the triangle they define. The symmetric tensor representing this deformation everywhere in the interior may be visualized directly by its effect on a circle, Figure 3b.

By changing the scale of one triangle or the other, we are free to superimpose the pair of triangles on any pair of vertices, for instance, A and B. In doing so, Figure 3c, we have of course altered the size component of the deformation relating the triangles; but the directions of the principal axes and (if the size difference is small) the anisotropy of the principal strains are left nearly unchanged. We use these rescaled registrations to extract the principal directions and extensions for small changes of form by the geometric construction shown in Figure 3d. Let the earlier and later positions C, C' of the third vertex be separated by the distance δ, and let h be the distance of vertex C from the baseline AB in the starting form.

Then the strains in the principal directions are approximately

$$d_1 = \frac{\delta}{h}\cos^2\alpha,$$

$$d_2 = -\frac{\delta}{h}\sin^2\alpha,$$

and so

$$d_1 - d_2 = \frac{\delta}{h}(\cos^2\alpha+\sin^2\alpha) = \frac{\delta}{h},$$

where α is the angle between the apparent displacement CC' and one principal axis, as drawn. We may also derive the relation

$$\alpha = 45° - \frac{\beta}{2},$$

where β is the angle between the path CC' and the fixed edge AB. This construction is explained in detail in Bookstein, 1984.

The apparent displacement of point C is made up of two vector components (Figure 3e): extension at a rate d_1' of the length L_1 along principal axis 1, and extension at a rate d_2' of the length L_2 along principal axis 2, where d_1', d_2' are the major and minor principal strains as scaled, that is, after the rate of extension along the baseline AB is subtracted. As long as we do not change the baseline, the ratio $L_1:L_2$ is the same for all choices of the point C. If the deformation is uniform, then, it will displace any other point D to its image D' by a vector DD' parallel to CC'. The vector DD' will be lengthened or shortened with respect to CC' according to the distance from D to the same baseline. (The position of D along line AB is irrelevant, as whenever a linear transformation leaves two points of a line

fixed it leaves all points of that line fixed.)

Interpreting the locations of points A, B, C as complex numbers, we can write this construction as the replacement of point C by $\left[\frac{C-A}{B-A}\right]$. (Cyclic permutation of A, B, C, referring point A, for instance, to the baseline BC rather than C to AB, replaces this ratio by one of the associated cross-ratios

$$\left[\frac{B-C}{A-C}\right] = 1 - \frac{1}{\left[\frac{C-A}{B-A}\right]}, \quad \left[\frac{A-B}{C-B}\right] = \frac{1}{1-\left[\frac{C-A}{B-A}\right]}.$$

We use this formalism to explore the consequence of perturbing two vertices for the displacement reported at a third. Suppose that landmarks A, B of ΔABC have shifted by vectors $[x_A, y_A]$, $[x_B, y_B]$. Even if vertex C of ΔABC is fixed in this coordinate system, Figure 4a, we can interpret the shape component of the deformation of ΔABC as a displacement imputed to C after registering on A and B as if they had been fixed instead. To simplify our algebra, let the reference coordinate system place point A at [0,0] before its shift, and point B at [1,0]. Point A has thus been shifted to point $A' = [x_A, y_A]$, and B to $B' = [1+x_B, y_B]$. Assume C is fixed at the point $[r,s]$ in this coordinate system; we wish to compute the displacement it undergoes when we register A' at A and B' at B. We will assume that the displacements AA', BB' are both small, so that we can ignore all second and higher powers of the x's and y's.

We have

$$\Delta C \sim \frac{\partial C}{\partial A}\Delta A + \frac{\partial C}{\partial B}\Delta B$$

$$= \frac{[(C-B)\,\Delta A - (C-A)\,\Delta B]}{(B-A)^2}.$$

Reverting to a vector notation, Figure 4b, the contribution of $[x_A, y_A]$, the variation at A, is scaled and rotated by the transformation taking [1,0] to $C-B = [r-1, s]$; the contribution of $[x_B, y_B]$, the variation at B, is scaled and rotated by the transformation taking [1,0] to $A-C = [-r, -s]$. For small changes, the factor B-A is almost constant and can be dropped. Then we have, for the coordinates of the normalized point C,

$$\begin{aligned} x_C &= x_A(r-1) + y_A(-s) + x_B(-r) + y_B(s), \\ y_C &= x_A(s) + y_A(r-1) + x_B(-s) + y_B(-r). \end{aligned} \qquad [1]$$

III. Vector Components of Shape Change for a Quadrilateral

This section shows how to generalize from triangles to quadrilaterals the interpretation of shape change through displacement: Instead of displacing vertices one at a time, we take them two at a time.

Superposition on the diagonal. It is convenient to begin with a starting configuration of landmarks that is exactly square. The diagonal of a square divides it into two triangles whose vertices are endpoints of the other diagonal. The shape change of the configuration of four points in a square may be represented by the simultaneous displacement of <u>both</u> these vertices when the other diagonal is fixed at both ends, as in

Figure 5a. The pair of vectors E and G together represent the shape change with reference to the particular diagonal we have selected.

We have specified one diagonal as "fixed," responsible for the displacements of the remaining two vertices. We might as well have reversed the roles of the two diagonals, Figure 5b. Equations [1] permit us to compute the vectors F, H which would have resulted had we done so. For the upper triangle, we use $[r,s] = \left[\frac{1}{2}, \frac{1}{2}\right]$; for the lower, $\left[\frac{1}{2}, -\frac{1}{2}\right]$. The components $[x_F, y_F]$, $[x_H, y_H]$ of F and H are thereby set forth in terms of the components $[x_E, y_E]$, $[x_G, y_G]$ of E and G:

$$\begin{aligned} x_F &= -\frac{x_E}{2} - \frac{y_E}{2} - \frac{x_G}{2} + \frac{y_G}{2}, \\ y_F &= \frac{x_E}{2} - \frac{y_E}{2} - \frac{x_G}{2} - \frac{y_G}{2}, \\ x_H &= -\frac{x_E}{2} + \frac{y_E}{2} - \frac{x_G}{2} - \frac{y_G}{2}, \\ y_H &= -\frac{x_E}{2} - \frac{y_E}{2} + \frac{x_G}{2} - \frac{y_G}{2}. \end{aligned} \qquad [2]$$

In this way we represent the same shape change twice over, using eight coordinates (four two-vectors E, F, G, H) instead of the mere four coordinates mathematically required.

Some identities. Manipulation of these equations leads to several interesting observations. We note first that, by virtue of cancellation of signs,

$$\begin{aligned} x_F + x_H &= -(x_E + x_G), \\ y_F + y_H &= -(y_E + y_G). \end{aligned}$$

In other words,

$$E + F + G + H = 0.$$

That is, the vector sum of the four imputed displacements, two each corresponding to the fixing of each of the two diagonals, is exactly zero. The diagram of four vectors thus bears no net shift component (Figure 5c).

We may also verify, by a different cancellation, that

$$\begin{aligned} x_F - x_H &= y_G - y_E, \\ y_F - y_H &= x_E - x_G. \end{aligned}$$

This may be rewritten in terms of Cartesian coordinates at 45° to the set we are presently using:

$$\begin{aligned} (y_E + x_E) + (x_F - y_F) + (-y_G - x_G) + (-x_H + y_H) &= 0, \\ (y_E - x_E) + (x_F + y_F) + (-y_G + x_G) + (-x_H - y_H) &= 0. \end{aligned}$$

That is, the "net component" of change representing expansion or contraction about the center of the square is exactly zero, and the "net component" representing rotation about the center is likewise exactly zero.

Two components. There are two other symmetries to be extracted from this coupled set of vector equations. Suppose, for instance, that $E = G = [x_{EG}, y_{FG}]$, as in Figure 6a. By substituting in equations [2], we find

$$x_F = x_H = -x_{EG} ,$$
$$y_F = y_H = -y_{EG} .$$

That is, if two opposite vertices are identically translated with respect to the other diagonal, then the endpoints of that diagonal show an equal and opposite translation with respect to the diagonal joining the first pair. The situation is therefore a motion of either diagonal with respect to the other without change of angle. We will call this case *pure translation* or the *purely inhomogeneous transformation*; it will occupy our attention throughout the remainder of this essay.

Instead of specifying E = G, we might instead consider the case E = -G: opposite corners of the square displaced equally and oppositely after registration upon the other pair of vertices. Substituting $x_E = -x_G$, $y_E = -y_G$ in equations (2), we obtain

$$x_F = -y_E ,$$
$$y_F = x_E ,$$
$$x_H = y_E ,$$
$$y_H = -x_E .$$

Hence F is a 90° clockwise rotation of E (or 90° counterclockwise rotation of G) and H, which equals -F, is a 90° clockwise rotation of G (or 90° counterclockwise rotation of E), Figure 6b.

In the displacements representing these deformations, the origins of vectors H and F (or E and G) are at equal and opposite distances from their common diagonal of reference, and the displacements imputed to them by the deformation are likewise equal and opposite. It follows (recall Figure 3c) that the

deformations of the square with E•G = F•H = 0 represent *pure* (or *uniform*, or *homogeneous*) *shears*, those transformations which are the same on both sides of (either) diagonal. The characterizations of F and H as rotated versions of E or G merely express the effect of a 90° change in baseline direction without change of baseline length.

Now for any vectors E, G whatever, it is the case that

$$E = \tfrac{1}{2}(E+G) + \tfrac{1}{2}(E-G) ,$$

$$G = \tfrac{1}{2}(E+G) - \tfrac{1}{2}(E-G) .$$

In this way we decompose <u>any</u> small deformation of the square into the composite of two deformations, one a pure shear and the other a pure translation. If one corner is displaced by E and the other by G, the pure shear component of the pair displaces one corner by $\frac{1}{2}(E-G)$, the opposite corner by the opposite translation; the pure translation component displaces each of the pair of opposite corners by $\frac{1}{2}(E+G)$. By the identities relating E, G and F, H that were established above, the magnitudes of these two components are invariant under change of the choice of diagonal.

The general starting quadrilateral. Up to change of size, we have reduced all deformations of a square to four specific components: pure shear, pure translation, pure shift (which is identically zero), and pure rotation/expansion (also identically zero). The deformation of the general quadrilateral manifests the same four components. Let us, for instance, compute the

picture of a pure shear for the quadrilateral of Figure 7 (top). Suppose the northwest corner is displaced by a vector E in relation to the northeast-southwest diagonal. The vector expressing the effect of this same shear upon the southeast corner is antiparallel to E with a length given by the ratio of the signed distances of these two corners to the other diagonal. There results the vector G as drawn. Were we to register this same shear upon the northwest-southeast diagonal instead, we would have rotated the baseline clockwise by 60° (the angle between the diagonals), and thus rotated the apparent displacement vector by 60° counterclockwise with respect to the baseline. The vectors F, H representing displacements of the other two corners are aligned in this new direction, and their lengths are proportional to the distances of <u>these</u> corners from their baseline diagonal. We may verify all this by noting that the two quadrilaterals drawn in dashes in Figure 7 are exactly similar.

IV. The Purely Inhomogeneous Transformation as a Mapping

The discussion of triangular configurations in Part II showed how to pass between two modes of description: change as a deformation, described by a (symmetric) tensor, and change as a relative displacement of landmarks, described by a (baseline-dependent) vector. This equivalence was extended to the homogeneous component of quadrilateral shape transformations. Up to scale change, for instance, a homogeneous shear of the square might be interpreted as equal and opposite vectors of

displacement at the endpoints of just one diagonal.

The case of the purely inhomogeneous component, Figure 6a, is rather less familiar. Part III described it only in terms of a displacement: translation of the diagonals with respect to each other, without rotation. For explanation of biological form-change—morphogenetic theory or morphometrics—we must pass to the transformational point of view and model this pure translation of diagonals as generated by a smooth deformation relating the interiors.

Consider, then, the task of modeling the mapping of the square into the kite, Figure 8a. I will refer to the left end of the square, the one apparently increasing in size, as the *positive pole*, and the right end, decreasing in size, as the *negative pole*. This diagonal will be called the *polar axis*; the other, the *nonpolar axis*. Smooth maps of the quadrilateral to the kite ought to express the aspect of "translation" in a manner analogous to the little vectors E, G of Figure 6a: relative extension or compression of length along the polar axis. Lengths measured along the vertical diagonal will be, on the average, unchanged.

The deformations to be considered are, like the square and kite forms themselves, symmetric around the polar axis. Everywhere on this axis of symmetry, the principal cross describing the deformation by its derivative must be aligned with horizontal and vertical, the polar axis itself and its perpendicular. At the right, the horizontal compression is the minor principal strain, dominated by the vertical quasi-stasis;

at the left, the horizontal expansion is the major principal strain. This major strain has thus rotated by 90° as we pass 180° around the form (and by 180° as we pass 360° around the form—in analogy to the winding number of a vector field, this feature implies a singularity inside).

The rotation of this principal direction by 90° may be clockwise or counterclockwise around the top of the form, as sketched in Figure 8b. In one case, which will be typified by the projection mapping, the directions of greater principal strain at the top and bottom corners of the quadrilateral pass approximately through the negative pole of the form. In the other case, typified by the bilinear mapping, the positive dilatation lingers upon the positive pole of the form.

Analysis of maps from either class involves a single parameter a characterizing the magnitude of the relative translation of endpoints embodying the deformation. In terms of the square starting form, a may be imagined a convenient multiple of the shift in either endpoint of either diagonal, scaled to the edge-length of the square. For both the classes of maps used as standard forms belo , there is only one distortion and one biorthogonal grid for the entire family of transformations. The starting square cuts out a greater- or lesser-sized area, depending on the value of a, from this single grid.

Projection. The reader will recall that projections, a familiar class of mappings, take all straight lines of the plane into straight lines, are linear in the extended plane of homogeneous coordinates, etc. The equations for projecting the

square onto the kite have a satisfactory symmetry when the square is positioned and scaled as in Figure 9a, with diagonals of length $2a$ centered at the point $[1,0]$ and aligned horizontally and vertically. The kite shares the vertical (nonpolar) diagonal through $[1,a]$ and $[1,-a]$—these points are fixed by the transformation—but its horizontal diagonal is shifted by the distance a^2, the fraction $a/2$ of diagonal length, at either end.

The (unique) projection which maps the four corners of this square onto the corresponding four corners of this kite has the equation

$$[x,y] \dashrightarrow \left[\frac{a^2-1}{x} + 2 - a^2,\ \frac{y}{x} \right] \qquad [3]$$

—one easily verifies that this mapping, a linear fractional transformation, leaves $[1,\pm a]$ fixed and takes $[1\pm a,0]$ to $[1\pm a-a^2,0]$. As drawn in Figure 9a, the map looks like an example of familiar Renaissance perspective.

The affine derivative of this projection at any point $[x,y]$ is given by the matrix

$$1/x^2 \begin{bmatrix} 1-a^2 & 0 \\ -y & x \end{bmatrix}$$

(operating from the left on column vectors). At the positive pole $[1-a,0]$ this matrix becomes $\mathrm{diag}\left[\frac{1+a}{1-a},\ \frac{1}{1-a} \right]$—expansion in all directions, but primarily horizontally. At the negative pole $[1+a,0]$ it becomes $\mathrm{diag}\left[\frac{1-a}{1+a},\ \frac{1}{1+a} \right]$—compression in all

directions, but primarily horizontally. The map thus exemplifies the qualitative problem which interests us, the rotation of the dominant principal strain by 90° as we pass 180° around the form, from the positive to the negative pole.

For applications to morphogenetic explanation we will find most interesting the behavior of this mapping near its *conformal point*, the point at which the derivative is mere combination of rotation and change of scale. At such points, even though the principal strain values are well-defined, they are equal; the principal axes do not exist, so that a crucial geometric parameter is undefined. A plane mapping has a conformal point wherever its affine derivative takes the form $\begin{bmatrix} c & -s \\ s & c \end{bmatrix}$, skew-symmetric with diagonal terms equal. For the projection, this is the case only at the point $[1-a^2,0]$.

The projection [3] of square into kite was derived by a combination of translation, reflection, and scaling from the canonical involutory projection $[x,y,z] \dashrightarrow [z,y,x]$, or, in Cartesian coordinates,

$$[x,y] \dashrightarrow \left[\frac{1}{x},\ \frac{y}{x} \right] .$$

In this version, $[1,0]$ is an *anticonformal point* at which angles are changed in sense but not in magnitude; the conformal point is at $[-1,0]$.

It may be verified by direct substitution that this map preserves the grid of confocal conics

$$\frac{x^2}{b} + \frac{y^2}{1+b} = 1$$

shown in Figure 9b. This grid of conics is at 90°; therefore it must itself be the biorthogonal grid pair (both coordinate meshes!) for the mapping. The action of the map is simply to interchange, for all $k>1$, the ellipse through $[k,0]$ with the hyperbola through $[1/k,0]$. The points on the y-axis of the ellipse go to the points at infinity on the matched hyperbola.

The line $x = 1$, the nonpolar axis, is pointwise fixed by the transformation; in particular, the center $[1,0]$ of the square is left fixed. The extension ratio along the polar axis is graded as x^{-2}. Nowhere except on the polar axis itself does either arm of any biorthogonal cross point to the conformal point. In particular, at the ends of the nonpolar axis, the principal strains are nearly at 45° to their orientations at the positive or negative poles. As in the left sketch of Figure 8b, the direction corresponding to the more positive strain points approximately to the negative pole, and vice versa.

It can be shown that the general projection of quadrilateral onto quadrilateral has exactly the same biorthogonal grid pair as this highly symmetrical example. In other words, up to rotation and change of scale, every projection of one quadrilateral onto another can be observed as the effect of this specific involution on some configuration of four starting corners. Every projection, whether the configuration of displacement it represents is pure shear, pure translation, or a combination, has a single conformal point and another (biologically irrelevant)

anticonformal point; and its grids are confocal conics about that pair of points. (Because projection leaves straight lines straight, all finite angles measured at the conformal point are unchanged, and all measured at the anticonformal point are merely reversed in sense; in this role the point has been useful to photogrammetrists for some time.) If the conformal point lies within the interiors of the forms, the grids have the aspect of this focus. If it lies outside, then everywhere one direction of gradient dominates, and the grid may be seen as a smooth warping of a rectangle, Figure 2c. These cases are located on the canonical confocal plot as in Figure 9c.

The bilinear mapping. Projection takes straight lines into straight lines but is highly nonlinear in its treatment of length along most lines. The homogeneous shear we seek to escape likewise takes straight lines to straight lines, and furthermore is linear along every line of the plane—which is why it lacks interest for the morphogeneticist. A compromise between these purposes is needed: a mapping that is linear on the edges of the quadrilateral, but that nevertheless allows for regional differences in directionality, in particular, for conformal points where the affine derivative is isotropic.

One class of maps satisfying these requirements is the *bilinear* family. While analytically almost as simple as the projection, they are nevertheless fairly unfamiliar to the applied geometer, and so I shall explain their algebraic and geometric properties in some detail.

The fundamental appearance of the bilinear mapping relating

a pair of quadrilaterals is as in Figure 10a. While projection deals with the quadrilateral as a set of four points in any order, bilinear mapping requires that one select four edges from the six possibilities. The transformation will be linear on these four edges, but not along the other two, those serving as the diagonals of the construction. Choose any point $[x,y]$ interior to this quadrilateral, and consider the set of all straight lines through it. Each line divides each edge of the quadrilateral in some ratio. We are interested in the lines through $[x,y]$ which divide one pair of opposite edges in the <u>same</u> ratio. For the general quadrilateral (no edges parallel) this is a quadratic criterion which results in a single pair of lines through $[x,y]$, lying on the point as drawn. For squares, the lines we seek are just the lines through $[x,y]$ parallel to the sides.

The image of the point $[x,y]$ under bilinear mapping is the intersection in the other form of the lines that divide the homologous edges in the same pair of ratios. In the Figure, for instance, line AA' divides edges P_1P_2 and P_3P_4 in the ratio 1:2, and line DD' divides both edges P_1P_3 and P_2P_4 in the ratio 1:3. We find the points which divide the edges of the opposite form in the same ratio: C (resp. C') divides Q_1Q_2 (resp. Q_3Q_4) in the ratio 1:2 and D (resp. D') divides Q_1Q_3 (resp. Q_2Q_4) in the ratio 1:3. Then the point $[x,y]$ at the intersection of AA' and DD' is mapped to the point at the intersection of CC' and DD'.

The analytic geometry of this construction becomes clearer when the quadrilateral $P_1P_2P_3P_4$ that we selected (that is, the

set of four edges chosen out of six) is circumscribed about a parabola, Figure 10b. The quadrilateral $Q_1Q_2Q_3Q_4$ may likewise be circumscribed about a parabola of its own. Because all parabolas are similar, up to a change of scale we may treat the polygon of P's and the polygon of Q's as circumscribed about the same parabola.

It is an old theorem that for any three tangents to a parabola, every other tangent is cut by the three in the same affine ratio. It follows that the lines we seek through any point $[x,y]$ are simply the unique pair of tangents to the parabola through $[x,y]$. If we assign each edge of the quadrilateral a coordinate by its intersection with any fixed tangent to the parabola (for example, the vertex tangent, Figure 10c), then the bilinear map is linear, separately, on the coordinates of the two tangents through $[x,y]$.

For starting forms which are exactly square, a different characterization of this map is available, one which is of some use in computer graphics: the bilinear map is the simplest *blending function* for the square. Suppose corners $[0,0]$, $[1,0]$, $[0,1]$, $[1,1]$ of a square are mapped into points Q_{00}, Q_{10}, Q_{01}, Q_{11} respectively. The vertical line through a point $[x,y]$ inside the square is the connection of the point $xQ_{11}+(1-x)Q_{01}$ on the top edge with the point $xQ_{10}+(1-x)Q_{00}$ on the bottom edge. The horizontal line through $[x,y]$ intersects this join at the point a fraction y of the way from bottom to top, the point

$$(1-y)\left[xQ_{10}+(1-x)Q_{00}\right] + y\left[xQ_{11}+(1-x)Q_{01}\right] .$$

Expanding, we see that the map sends $[x,y]$ to the point

$$(1-x)(1-y)Q_{00} + x(1-y)Q_{10} + (1-x)yQ_{01} + xyQ_{11}$$

—a weighted average of the four image points corresponding to the four corners of the square, each weight given by the product of the distances from $[x,y]$ to the pair of grid lines through the diagonally opposite corner of the square.

Our particular concern is the bilinear map corresponding to the purely inhomogeneous transformation (pure translation), square onto kite. The algebra of this map is simplest under a standardization somewhat different from that of Figure 9a, which simplified the algebra of projection. Regardless of the parameter a of the kite, place the corners of the starting square at $[\pm1,\pm1]$. The image quadrilateral for transformations having shift parameter a shifts the main diagonal (the polar axis) by the vector $[-2a,-2a]$, a fraction a of its length. Hence, while the ends of the nonpolar axis, $[-1,1]$ and $[1,-1]$, are left fixed, the point $[-1,-1]$ is mapped to $[-1-2a,-1-2a]$, and likewise $[1,1]$ to $[1-2a,1-2a]$. The bilinear mapping for this pair of quadrilaterals is shown in Figure 11a; its equation is

$$[x,y] \mapsto [x,y] - a[1+xy,1+xy] . \qquad [4]$$

The derivative of this map may be found by direct calculation or by bilinear mixing of its derivatives at the four corners; it may be written in matrix form as

$$\begin{bmatrix} 1-ay & -ax \\ -ay & 1-ax \end{bmatrix}.$$

The bilinear map is potentially biologically meaningful wherever the determinant $1-a(x+y)$ of this matrix is positive. Whenever a is less than 0.5, this will be the case throughout the square. For $a=0.5$, the two edges $Q_{11}Q_{01}$, $Q_{11}Q_{10}$ of the kite lie on the same line (see Figure 11e), so that the image Q_{11} of the point [1,1] is undefined. The projection map likewise had a line of singularities, the real axis; while for projection areas near the line "blow up," the bilinear map squashes them flat.

At [-1,-1] the derivative of the bilinear map is $\begin{bmatrix} 1+a & a \\ a & 1+a \end{bmatrix}$, representing an <u>increase</u> in size in all directions except the nonpolar axis; then the polar axis bears the major principal strain. At [1,1] the derivative $\begin{bmatrix} 1-a & -a \\ -a & 1-a \end{bmatrix}$ represents a <u>decrease</u> of size in all directions except the nonpolar axis; the polar axis here bears the minor principal strain. Thus the qualitative behavior of this map, like that of the standardized projection, suits the import of the purely inhomogeneous transformation: the major principal strain rotates by 90° as we travel halfway around the starting square.

Notice that along the polar axis, axis of symmetry of the scene, the bilinear mapping bears a different nonlinearity from that of the projection. The image of the point $[x,x]$ is $[x-a-ax^2, x-a-ax^2]$; the derivative of this map is linearly graded

from $1+2a$ to $1-2a$ along the polar axis, and the derivative perpendicular to that axis is constant at 1. The image of the center [0,0] of the square is not the intersection of the diagonals of the kite, but rather the point $[-a,-a]$ displaced from [0,0] by half as much as the endpoints of the major diagonal are displaced from <u>their</u> starting loci. Notice, too, that the minor diagonal is not left straight under this mapping; it is mapped (cf. Figure 11a) into the arc of a parabola.

The representation of this mapping by the symmetric tensor field corresponding to its derivative is shown in Figure 11b. As we pass over the upper half of the square from [1,1] to [-1,-1], the major principal strain for the bilinear map rotates in a sense opposite to that for the projection. At the ends of the nonpolar axis, the principal strains are again at about 45° to those at the endpoints of the polar axis; but the greater principal strain points toward the positive, not the negative, pole. This is associated, naturally, with the linearity the bilinear map enforces on each edge of the quadrilateral separately.)

Grids near the conformal point of the bilinear map. There is a conformal point located where $1-ay = 1-ax$, $ax = -ay$, i.e. [0,0]. The behavior of the biorthogonal grid system around the conformal point may be classified (Bookstein, 1978, pp. 103-107) by studying the lines through the conformal point that are themselves included in the biorthogonal grid. For projections, there is only one such line, the polar axis itself. For the bilinear map, we can see that there must be two others.

Figure 11c shows a few positions of the principal cross along a circular path from positive pole to negative pole. At the positive pole, the major strain axis passes through the conformal point. As both strains are rotating clockwise along this counterclockwise circuit, soon the <u>minor</u> strain axis must sweep through the conformal point; and shortly afterward the <u>major</u> strain must pass through that point again; and, finally, the minor strain again—but by then we have arrived at the negative pole of the form.

This intuitive deduction can be verified analytically without much difficulty. One of the principal strains will pass through the conformal point whenever the affine derivative map at $[x,y]$ leaves orthogonal the line-elements $[x,y]$ and $[-y,x]$ upon the square. The images of these directions upon the kite are

$$\begin{bmatrix} 1-ay & -ax \\ -ay & 1-ax \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x-2axy \\ y-2axy \end{bmatrix};$$

$$\begin{bmatrix} 1-ay & -ax \\ -ay & 1-ax \end{bmatrix} \begin{bmatrix} -y \\ x \end{bmatrix} = \begin{bmatrix} -y-ax^2+ay^2 \\ x-ax^2+ay^2 \end{bmatrix}.$$

The right-hand sides are perpendicular whenever their dot product vanishes; there results a single equation in x and y,

$$(4ax+1)y^3 + (3a)y^2 - (4ax^3+3x^2)y - x^3 = 0.$$

One root of this equation is $x = y$, the polar axis. The other two roots satisfy

$$(4ax+1)y^2 + (4ax+4)xy + x^2 = 0$$

or

$$(4ax+1)\lambda^2 + (4ax+4)\lambda + 1 = 0$$

where $\lambda = y/x$, the slope of the direction out of the conformal point.

Near $[x,y] = [0,0]$, the equation reduces to $\lambda^2 + 4\lambda + 1 = 0$ with roots $-2\pm\sqrt{3}$. But these values are the tangent and cotangent of 105°. These roots are at exactly ±60° to the root $\lambda=1$ already located along the polar axis, so that near our conformal point the biorthogonal grids have perfect hexagonal symmetry. As ax moves away from 0, the orientation of these roots rotates slightly. All this is visible in the empirically computed biorthogonal grids for this bilinear mapping, Figure 11d.

The general bilinear map of the square, whether or not purely inhomogeneous, corresponds to the map of some starting square according to the Cartesian grid of Figure 11e and common biorthogonal grid of Figure 11f. Purely inhomogeneous maps of various parameters a correspond to squares of various sizes about the common conformal point [0,0] of the mapping; the smaller squares have smaller shift parameters. Transforms with a homogeneous component shift the starting square away from the conformal point [0,0]; if the shift is great enough, the conformal point will pass outside the boundaries of the starting square, and the grids that result will have the simpler topology of the distorted rectangle with one polarity always dominant, Figure 2c. For the general bilinear mapping, which does not begin with a starting form precisely square, the form of the grids near the singularity may be shown to be exactly the same.

Away from the singularity, there is an additional twisting, or spiral turning of the system.

The generic difference between these mappings. The qualitative distinction between these two classes of grid behavior around the conformal point is stable against changes of form of either the starting or the finishing quadrilateral and stable, also, against changes of detail in the interpolated homology map. When the direction of the major principal strain corotates in the passage from the positive pole to the negative pole of the form, Figure 8b left, the conformal point bears a singularity with a single strain trajectory through it. In its vicinity, the grid system resembles a confocal system of ellipses and hyperbolas near one focus—which, in turn, resembles the parabolic coordinates of classical mathematical physics, Figure 1b. When the major principal strain counterrotates as we pass from the positive to the negative pole, Figure 8b right, the conformal point supports three strain trajectories. (For the projection the additional two were imaginary: the isotropic lines $1 = y/x = \pm\sqrt{-1}$.)

To these different senses of the rotation of the principal major strain correspond different behaviors of the mapping along the boundary. The bilinear mapping is uniformly linear on each edge separately; the projection mapping is nonlinear along each edge, graded from positive to negative as one passes away from the positive pole and from negative to positive as one passes away from the negative pole.

V. Tensor Fields and Morphogenetic Explanation: Some Speculations

Putative morphogens based on reaction-diffusion models are relative chemical concentrations, scalar fields having a well-determined value or values, in principle, at every point of a region. The only singularities they are likely to have are due to zeroes in the denominators of ratios. Gradients, along with related concepts such as phase, are vector fields having a direction and magnitude at every point; singularities arise where the magnitude is zero, for there the direction is undefined. Symmetric tensor fields, as depicted by a biorthogonal grid, bear two perpendicular directions and two magnitudes at every point, and their singularities are of a different topology than those of the lower-order fields. While at the singularity of polar coordinates no direction is defined, at the singularity of a symmetric tensor field we see preferred directions defined in a geometrically new way. While polar coordinates have only the one topological form of singularity, the symmetric tensor field has two different forms: one with a single preferred direction, the other with three.

I believe it is time for a major rethinking of the whole notion of morphogenetic field. Current metaphors, such as are used in the discussions of compartment boundary formation elsewhere in this volume [Editor—true?], tacitly assume morphogens which are either scalar or vector fields. Discussions of singularities are thereby impoverished and explanations of experimental findings made defective to an unknown extent. For

tensors are as real as scalars and vectors: a cell or tissue has mechanical integrity, after all, so that a structurally encoded orientation could as easily be a strain axis (tensor) as a gradient (vector) or chemical concentration (scalar).

I have no evidence in hand that this possibility is in fact the case—I know of no attempts to measure a real tensor field and match its singularities to organizing centers for ensuing morphogenetic phenomena. But I would nevertheless draw the reader's attention to four possible applications of tensors in morphogenetic explanation: (1) the symmetrical bifurcation of axial gradients of growth; (2) the orientation and anatomical meaning of certain shape-change grids observed empirically; (3) the emergence of the three-root singularity from juxtaposition of two one-root singularities; and (4) the emergence of reorganization from a one-parameter model mixing homogeneous and inhomogeneous components of change.

1. Bifurcation of axial gradients. Conventional compartment models locate axes by resonances or catastrophes within the interior shape; axes modeled to have no natural polarity. But in the study of growing systems, which are subject at all times to real, material strains, the model of singularities of tensor fields may be much more appropriate to an explanation of the physical branching structures which result. For instance, in regeneration of the newt limb (Bookstein and Connelly, in prep.) a longitudinal axis bifurcates, Figure 12a, as successive digits appear. The resulting bifurcated fields still point "forward"; the original axial polarity is maintained

in spite of the bifurcation. This phenomenon suggests a model bearing two channels of information, one for direction and one for spatial gradient independent of direction; the model of a strain trajectory seems well-suited, as this construct may have a singularity of direction while its numerical magnitude is behaving quite properly. The bifurcation of a pair of digits, then, might be viewed as the splitting of the positive-right trajectory of Figure 12b in the vicinity of the parabolic singularity.

An alternate model for this same bifurcation is available using the other generic singularity, the one with triple symmetry. Under the bilinear mapping, two adjacent edges of the boundary grow relatively more than the other two—in effect, grow over the other two while the width of the system is maintained constant. This resembles the actual situation in the regenerating newt limb: while the limb maintains a fairly constant width after amputation, and actively increases in length, a cap of epithelium at the end, the *apical cap*, soon ceases to grow. In the literature of this phenomenon, the cap has been considered an inducer of differentiation slightly proximal to itself. But it could as well induce the bifurcating tensor field of Figure 12c, so that condensing protocartilage finds itself with multiple principal directions rather than the previous single axial orientation.

2. The singularity of human calvarial growth. In the normal growth of the human cranium viewed laterally, Figure 13, there is an apparent singularity in the vicinity of the sella

turcica (pituitary fossa)—an apparent "growth center." The same pattern is seen in comparisons of juvenile with adult shapes throughout the anthropoid primates (Bookstein, 1978). Using another mathematical model, Moss et al. (1981) located the same center for the head as a so-called *allometric center*, a sort of mean conformal point based in measurement of angles at a distance rather than in the local structure of a tensor field. The same authors later rejected this notion (Moss et al., 1983), in view of its statistical imprecision. In my view the retraction was in error. The locus they extracted was of morphogenetic significance; however, their mathematical model lacked geometrical language for discriminating the tensor interpretation from the vector.

The traditional explanation of this general change in form is the positive allometry of the jaws with respect to the brain. Notice, however, that the invariant axis of the singularity is not oriented anteroposteriorly, along that gradient, at all. The shape change along this axis bears a substantial component of shear, owing to the "orthocephalization" of cranial growth, the movement of the jaws forward, out from under the brain-case. The principal strain trajectory through the singularity instead runs more or less vertically, from a point in the forehead down to the insertion of the spinal cord. This suggests additional, rather different explanations of the regulation of the change, explanations perhaps associated with maintaining the orientation of the visual system or the mechanical advantage of the muscles of mastication. In any case, the detailed geometry of

deformation of the cranium carries more information than a mere scalar summary of relative rates of increase of a size measure in the various regions.

3. Emergence of a three-root singularity from the juxtaposition of two one-root singularities. Figure 14a models the two basic singularities of the tensor field, that from the projection and that from the bilinear mapping, using a notion of "centers" under conditions of horizontal polarity and vertical constraint. The location of the singularity is at the open circle, the "constraint" (which is merely our measure of scale) by arrows to the two landmarks held at (relatively) constant spacing. We may then sketch the principal axes of the grid corresponding to the projection as a system of polarities radiating approximately from the singularity at the center. The axes of the grid corresponding to the bilinear mapping, on the other hand, appear to radiate instead from the positive and negative poles separately.

When we juxtapose two of these parabolic organizing fields, as in Figure 14b, in-between the two attracting centers emerges the aspect of the bilinear field (with its direction reversed). There develop three positive gradients rather than two. In other words, *from the abutting of two fields, each with a single privileged direction, there emerges a new field with three.* Figure 14c demonstrates this explicitly, averaging the two centered maps by the elastic algorithm of Bookstein (1978).

There is a classic set of morphogenetic experiments to which this observation is directly relevant. When one abuts two same-

sided embryonic chick wings on the same side of the embryo, or graft two newt limb stumps to the same regeneration site, there are often observed not <u>two</u> wings in the adult animal but <u>three</u>, and the middle one is reversed in polarity (a left wing between two right wings, or vice versa). These experiments have long been used to support the model of intercalation of polar coordinates. But on the tensor model, the appropriate kinematic model is not intercalation of a coordinate, but irreversible bifurcation of a polarity.

If a full axis of these parabolic polar centers is established (perhaps as the result of some resonance process—I am not a tensor chauvinist), the summated morphogenetic field will induce a symmetrical pair of positive gradients at 120° for every initial center. This seems to me a most suggestive metaphor for the production of bilateral structures by segmentation, a process for which there is currently no satisfactory mathematical model.

4. Timing of reorganization as admixture of components. I conclude this set of speculations by recalling from Part III the two components, homogeneous and inhomogeneous, of quadrilateral transformations. Under either bilinear mapping or projection, the purely inhomogeneous transform of a square manifested a conformal point precisely at its center. Displacement of the conformal point is equivalent to augmenting the transformation by a homogeneous component. When that component is sufficiently strong, the conformal point is displaced quite outside the boundary of the region we are studying, so that the influence of

inhomogeneity is seen only in the gentle curving of the grid lines, Figure 2c, and the shallow gradients along them: nonlinearity rather than polarity.

I suggest a time-dependent admixture of this homogeneity as an alternative to current explanations of crucial staging events in morphogenesis. Catastrophe models presume a scalar parameter that is varied past the point at which solutions of a variation equation change their topological properties. In the tensor model, the critical points represent not bifurcation of global minimizing criteria but spatial competition of gradient patterns. Bifurcation can arise directly from the sum of two tensor fields, each constant over time, if the ratio of their relative strengths varies. Biological processes often proceed by independent time scales, suggesting a likely site for a putative "master morphogen" regulating the combination. Variation of the mixture of fields over time is equivalent to a drift of the conformal point in geometrical space. When it encounters tissue, then and there geometric ramifications of structure may begin. This model is at least as congenial as those requiring nonlinear regulation of the entire system for generation of global critical points, resonances, and the like.

Acknowledgement. The writing of this essay was supported by N.I.H. grants DE-05410 to Fred L. Bookstein and DE-03610 to Robert E. Moyers. The Fortran program BIORTH which produced the symmetric tensor fields and the biorthogonal grids summarizing them was underwritten by the preceding grants and also by N.S.F. grant SOC 77-21132 to Fred L. Bookstein. Part V benefited by

conversations with Thomas Connelly of the Department of Anatomy at the University of Michigan.

Literature Cited

Bôcher, Maxime. 1894. *Über die Reihenentwicklungen der Potentialtheorie*. Leipzig: Teubner.

Bookstein, Fred L. 1978. *The Measurement of Biological Shape and Shape Change*. Berlin: Springer-Verlag.

Bookstein, Fred L. 1981. Coordinate systems and morphogenesis. Pp. 262-282 in *Morphogenesis and Pattern Formation*, ed. T.G. Connelly et al. New York: Raven Press.

Bookstein, Fred L. 1984. A statistical method for biological shape change. *Journal of Theoretical Biology*, in press.

Bookstein, Fred L., D. Chernoff, R. Elder, J. Humphries, G. Smith, and R. Strauss. *Shape Comparisons in Fishes: an Introduction to Morphometrics*. Completed manuscript.

Jacobson, A. G., and R. Gordon. 1976. Changes in the shape of the developing vertebrate nervous system analyzed experimentally, mathematically, and by computer simulation. *Journal of Experimental Zoology* 197:191-246.

Moss, M. L., R. Skalak, L. Moss-Salentijn, G. Dasgupta, M. Vilmann, and P. Mehta. 1981. The allometric center. The biological basis of an analytical model of growth. *Proceedings of the Finnish Dental Society* 77:119-128.

Moss, M. L., R. Skalak, M. Shinozuka, L. Moss-Salentijn, and M. Vilmann. 1983. Statistical testing of an allometric centered model of craniofacial growth. *American Journal of Orthodontics* 83:5-18.

Schüepp, O. 1966. *Korlstone*. Basel: Birkhauser Verlag.

Thompson, D'A. W. 1961 (1917, 1942). *On Growth and Form*, ed. abr. J. T. Bonner. Cambridge: the University Press.

Winfree, A. T. 1980. *The Geometry of Biological Time*. New York: Springer Verlag.

Captions for Figures

Figure 1. Two garden-variety oddities of orthogonal coordinate systems. (See text.) (a) A bicircular quartic coordinate system. (From Bocher, 1894.) (b) Parabolic coordinates.

Figure 2. Representation of a homology map by a symmetric tensor field. (a) The "Cartesian grid," after Thompson (1961), smoothly interpolating the correspondence of corners around the boundary and throughout the interiors of the forms. (b) The tensor field for the mapping: its symmetrized affine derivative, point by point. The length of each cross arm is drawn proportional to the principal strain along it. (c) The biorthogonal grid pair, a collection of integral curves of the tensor field of (b). Intersections of curves are homologous, left to right, according to the map in (a), and are at 90° everywhere in both forms. A few principal strains are indicated, ratios of lengths right:left, corresponding to the segments of arc over which they lie.

Figure 3. Expression of deformation tensor by a displacement with respect to a baseline. (a) A homogeneous deformation, after Thompson. (b) The tensor representing this deformation, its axes the principal axes of the ellipse into which a circle is deformed. (c) Superposition of two triangles upon a common baseline. A change of scale is required. (d) Construction of the deformation tensor (up to a scale factor) from the displacement vector. (e) Construction of the displacement vector from the

tensor.

Figure 4. Propagation of displacement. (a) Points A and B are shifted in a coordinate system holding C fixed. (b) The effective shift of C when we register on the shifted A and B by rotation and rescaling. In the figure, the asterisk represents complex multiplication.

Figure 5. Displacements of corners of quadrilaterals. (a) Diagonal FH fixed, corners E and G displaced. (b) Diagonal EG fixed, corners F and H displaced.

Figure 6. The two components of shape displacement for squares. (a) The purely inhomogeneous transformation, translation without rotation of one diagonal with respect to the other. (b) Pure homogeneous shear.

Figure 7. Sketch of a pure shear component for a general starting quadrilateral.

Figure 8. Mapping the square into the kite. (a) Symmetries, poles, and axes mentioned in the text. (b) There are two possible senses of rotation of the minor principal strain over the path from the positive to the negative pole.

Figure 9. Projection of the square onto the kite. (a) Placement of forms for $m=.25$, and Cartesian grid of the mapping. (b) The conformal point and biorthogonal grid of confocal conics about it. (c) Every projection is duplicated, up to a change of scale, somewhere on this grid.

Figure 10. The bilinear mapping. (a) Through the point $[x,y]$ pass two lines dividing opposite edges of the quadrilateral in equal ratios. The image of $[x,y]$ is the intersection in

the other quadrilateral of joins of the points dividing their edges in the same ratios. (b) Quadrilaterals in general position can be circumscribed about a parabola. The lines through (x,y) in frame (a) are the two tangents through (x,y) to this parabola. (c) The bilinear map is linear, separately, on the coordinates of the intersections of these tangents with any fixed tangent to the parabola.

Figure 11. Bilinear mapping of the square onto the kite. (a) Placement of forms for $a=.25$, and Cartesian grid of the mapping. (b) The symmetric tensor field representing this map by its affine derivative. (c) Why there must be three lines through the conformal point that lie on the biorthogonal grid. (See text.) (d) The biorthogonal grids for the mapping in (a), with some principal strains indicated. (e) The full bilinear tableau for the entire half-plane (cf. Figure 9c), and its single biorthogonal grid. (f) In the mapping from square to quadrilateral, the presence of a homogeneous component displaces the conformal point from the center of the square. Changes in the value of a alter the scale at which this single grid bears the transformation in question.

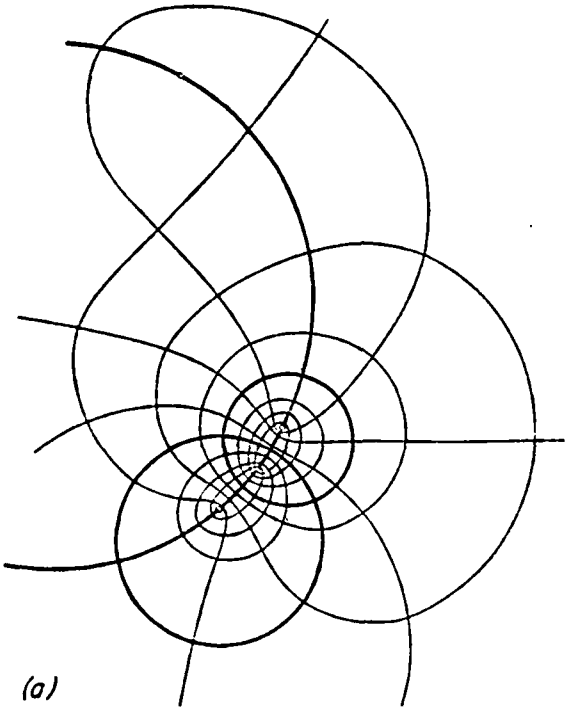
Figure 12. Bifurcation of the developing vertebrate limb skeleton. (a) Typical findings. (b) One model: jump to another branch of the projective singularity. (c) Another model: development along all axial polarities of the three-root singularity.

Figure 13. Biorthogonal grids for the observed growth of the

human calvarium. (From Bookstein, 1978, Fig. VII-10.) Note the one-root singularity.

Figure 14. Bifurcation of polarities from juxtaposition of singularities. (a) Sketch of the two basic singularities. (b) The structure of a three-root singularity appears between successive one-root centers. (c) Simulation of this bifurcation by relative displacement of eight homologous points. Left, the mapping; right, its grids.

Figure 1a



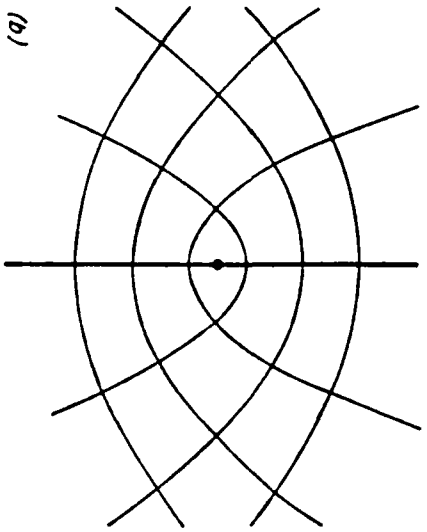


Figure 1b

Figure 2a

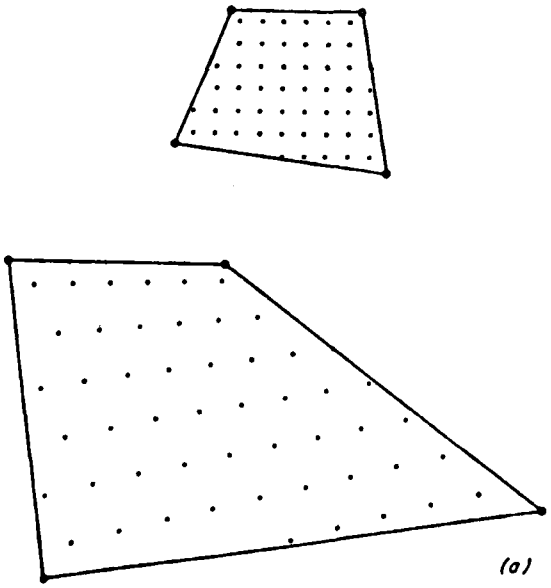


Figure 2b

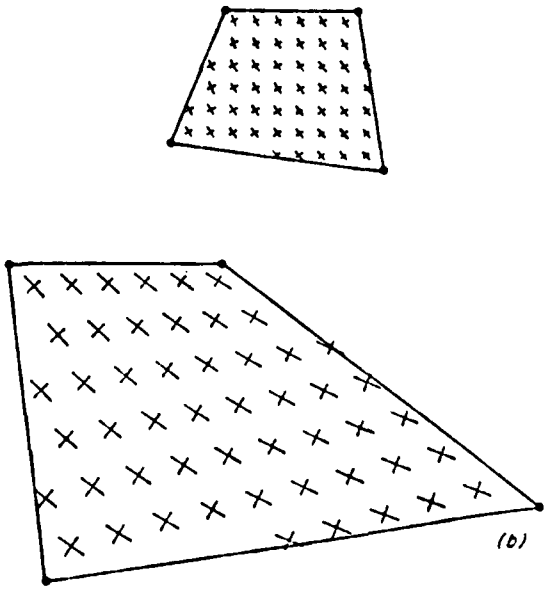


Figure 2c

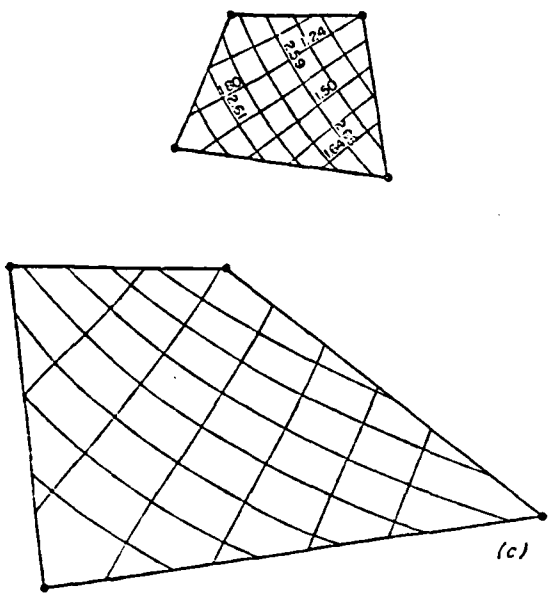


Figure 3a-c

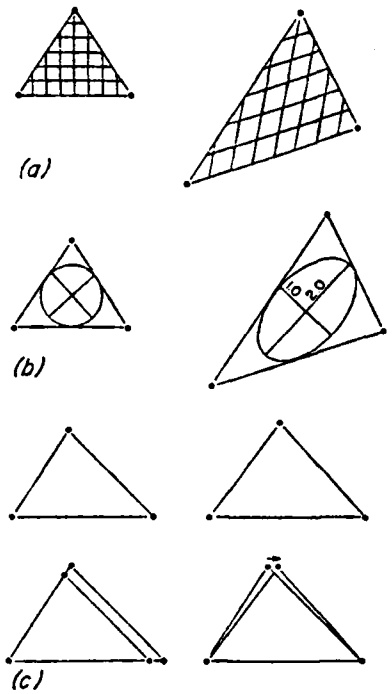
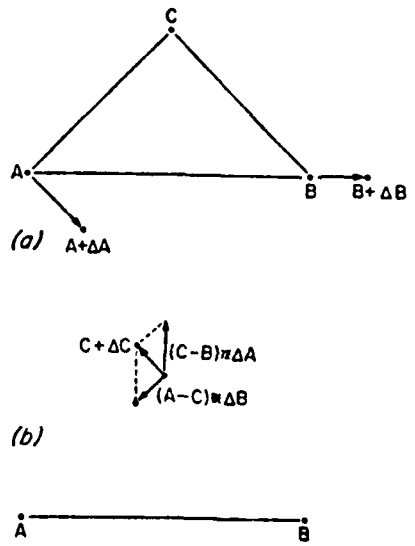


Figure 4



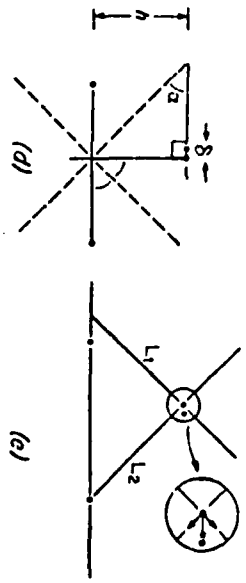


Figure 5a-e

Figure 5a-b

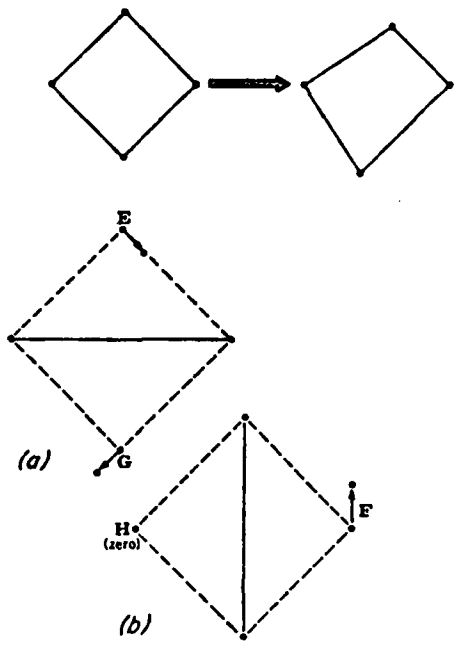


Figure 5c

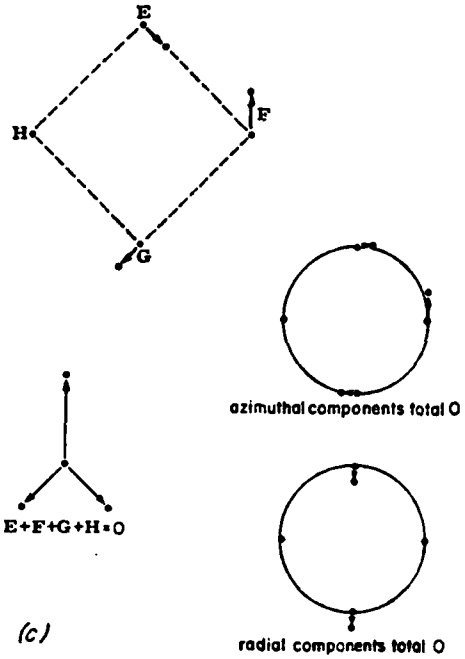


Figure 6

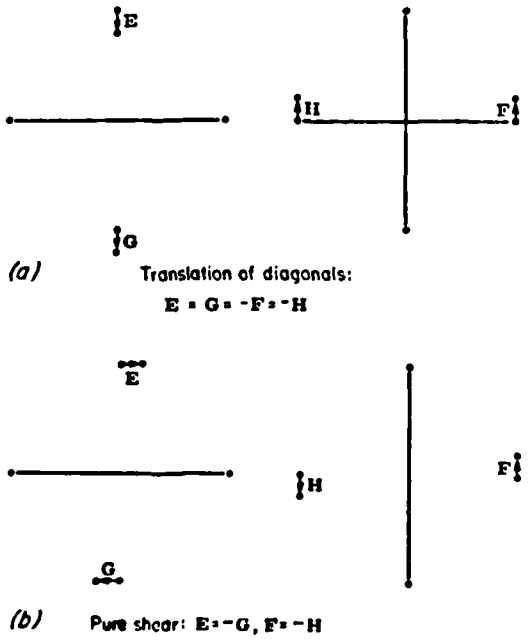


Figure 7

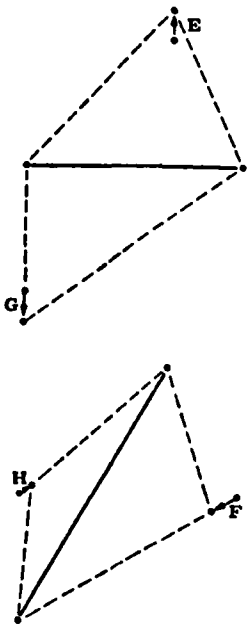


Figure 9a

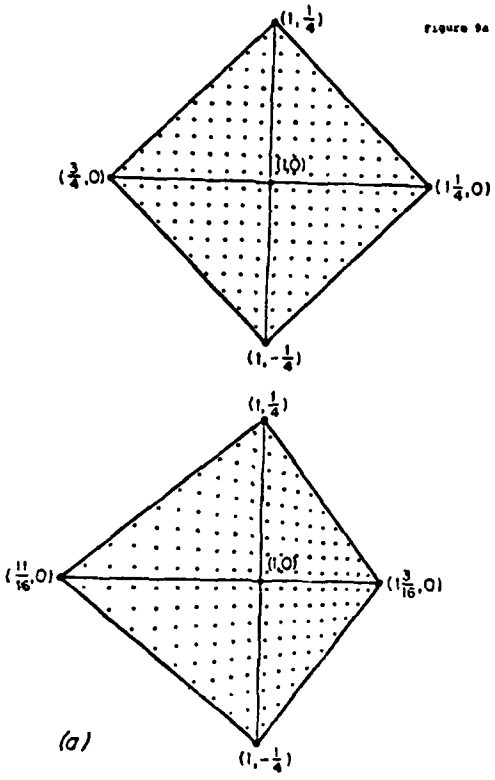


Figure 8

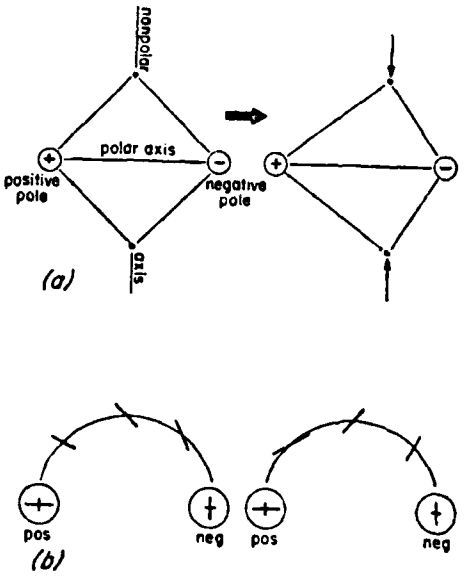
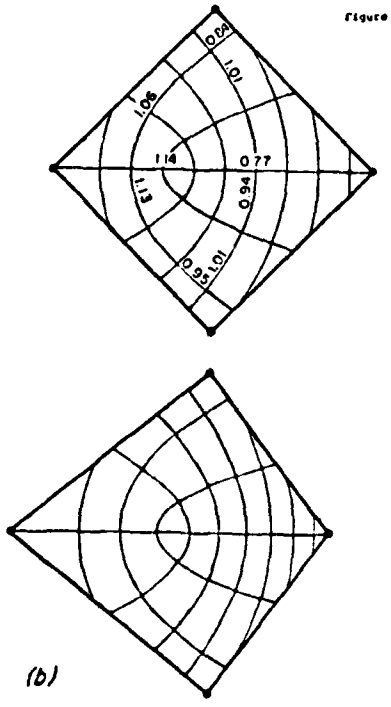
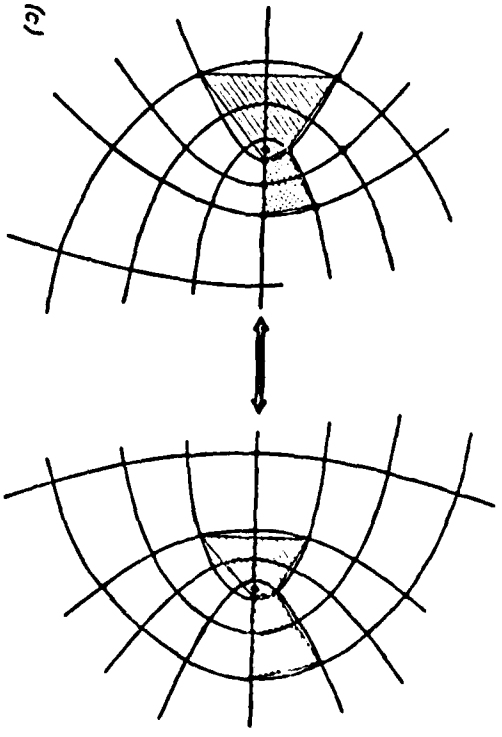


Figure 9b





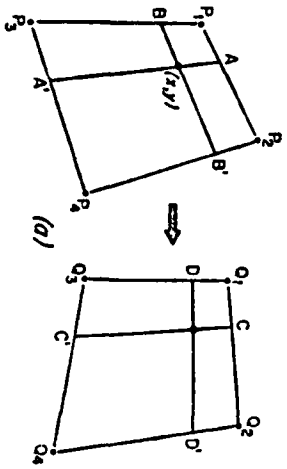


Figure 10a

Figure 10b

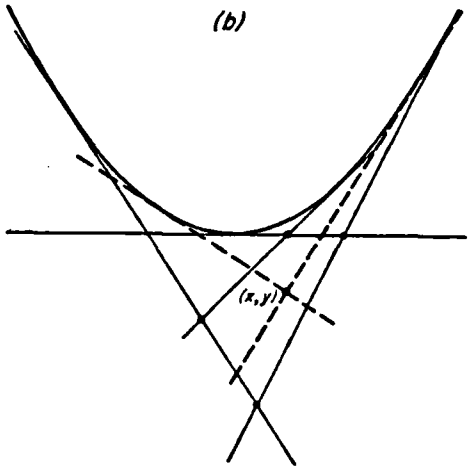


Figure 10c

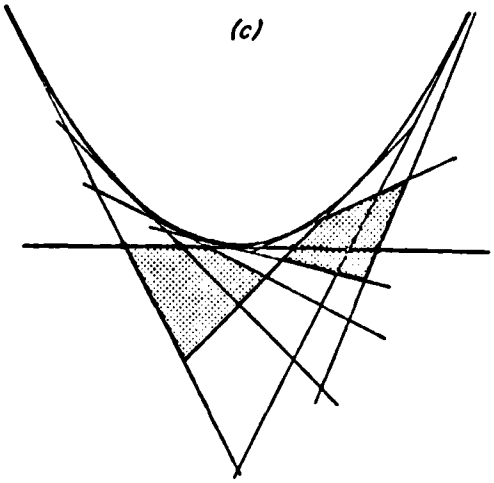


Figure 11a

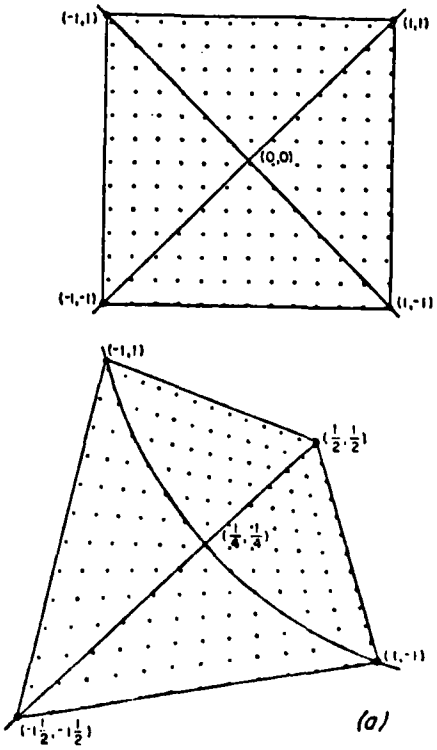


Figure 11b

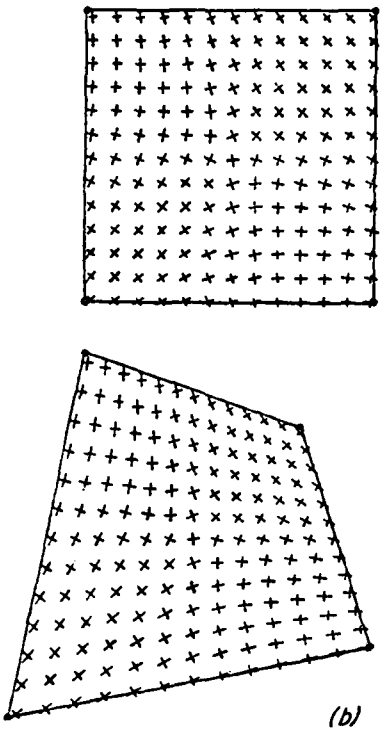


Figure 11c

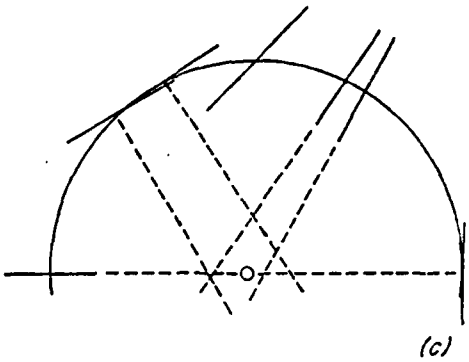
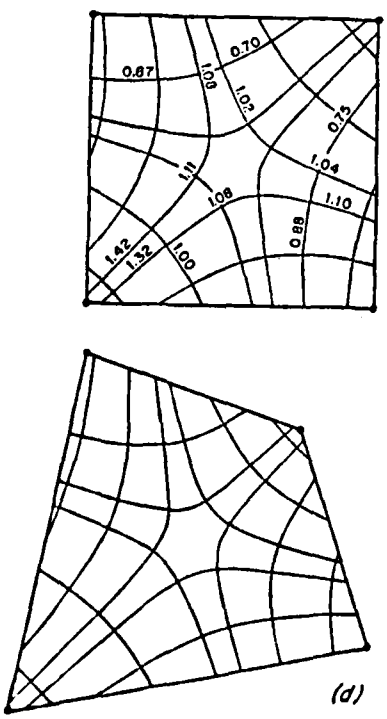


Figure 11d



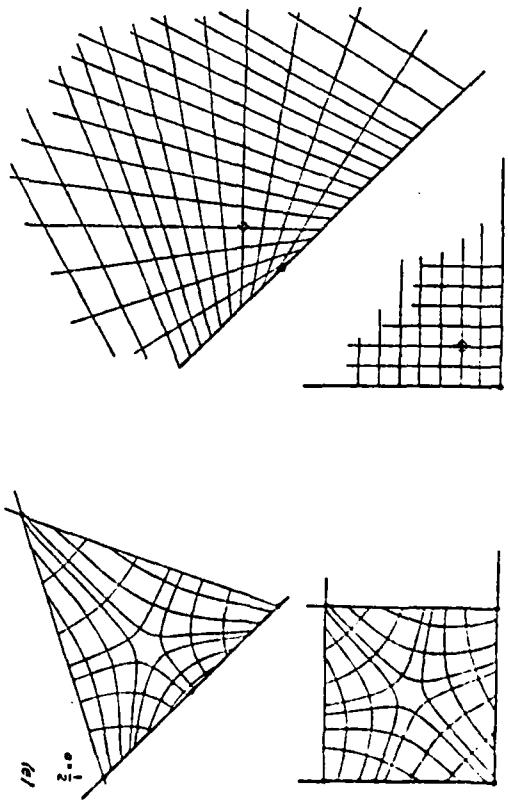


Figure 11f

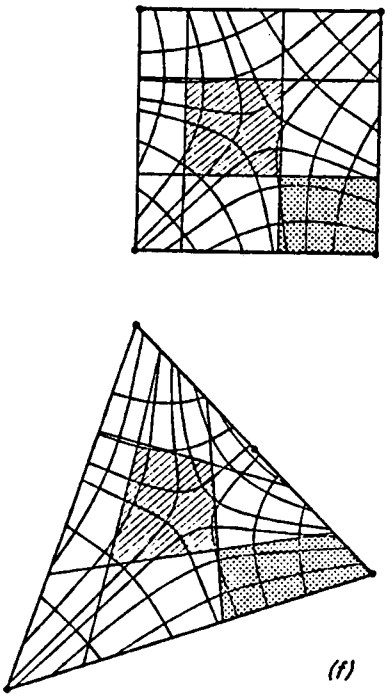


Figure 12a

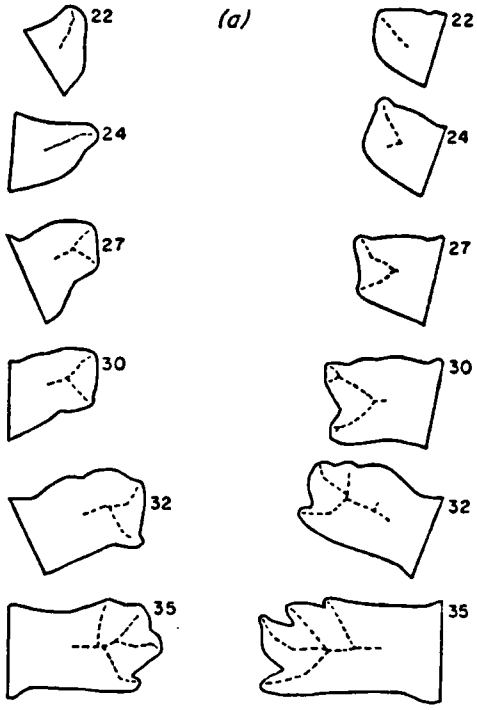


Figure 12b-c

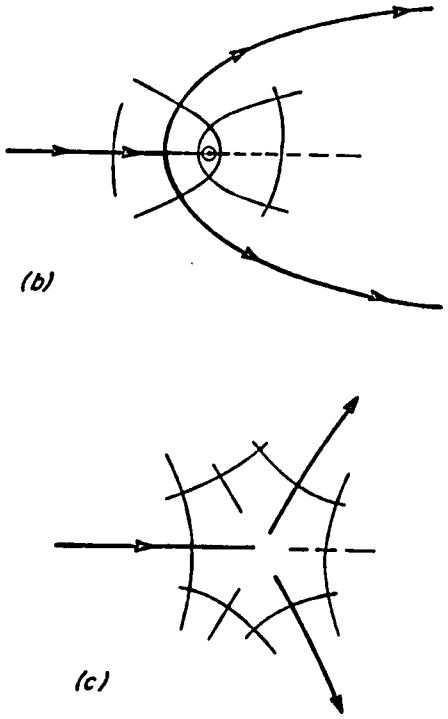


Figure 13

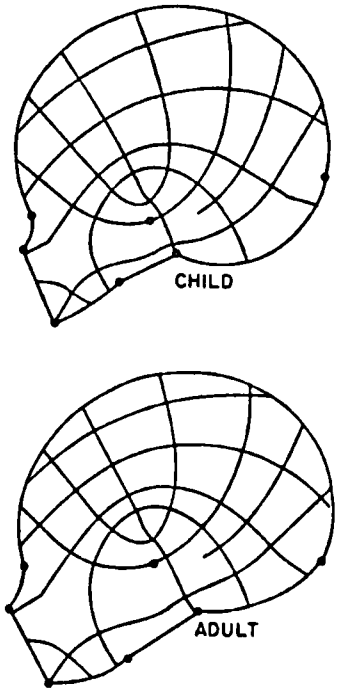
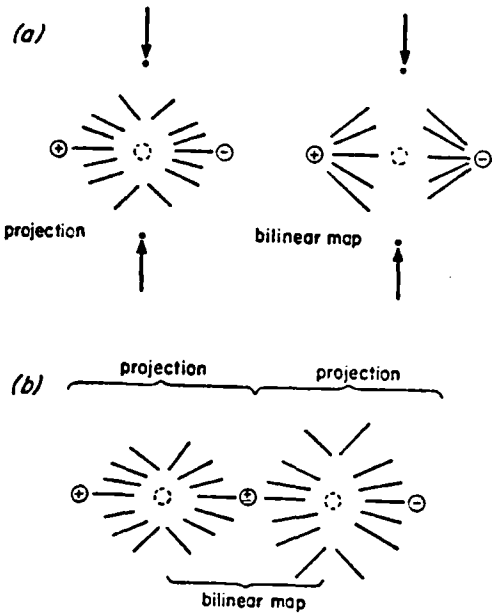


Figure 14a-b



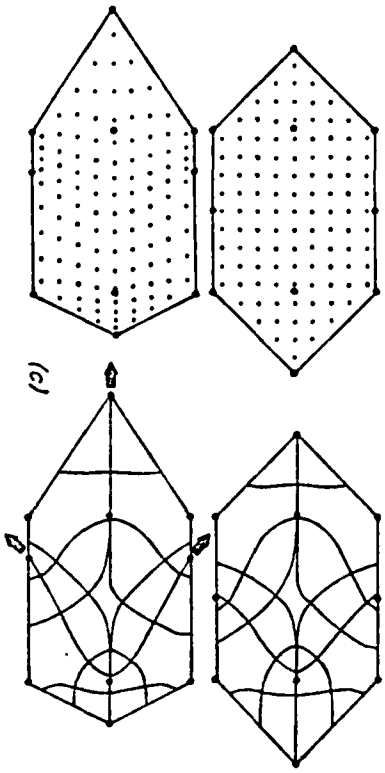


Figure 14c

SURFACES IN COMPUTER AIDED GEOMETRIC DESIGN:

A Survey with New Results

R. E. Barnhill
Department of Mathematics
University of Utah
Salt Lake City, Utah 84112

Abstract

"Surfaces in Computer Aided Geometric Design" focuses on the representation and design of surfaces in a computer graphics environment. This new area has the dual attractions of interesting research problems and important applications. The subject can be approached from two points of view: The design of surfaces which includes the interactive modification of geometric information and the representation of surfaces for which the geometric information is relatively fixed. Design takes place in 3-space whereas representation can be higher dimensional. "Surfaces in CAGD" can be traced from its inception in rectangular Coons patches and Bezier patches to triangular patches which are current research topics. Triangular patches can interpolate and approximate to arbitrarily located data and require the preprocessing steps of triangulation and derivative estimation. New contouring methods have been found using these triangular patches. Finally, multidimensional interpolation schemes have been based on tetrahedral interpolants and are illustrated by surfaces in 4-space by means of color computer graphics.

Running title: Surfaces in CAGD

Key words: Surfaces, Interpolation, Approximation, Design of surfaces, Representation of surfaces, Coons patches, Bezier patches, Triangular patches, Contouring, Multidimensional surfaces, Computer graphics

To appear in Surfaces in Computer Aided Geometric Design '84, R. E. Barnhill and W. Boehm, editors, North-Holland (1985) and in the journa omputer Aided Geometric Design.

TABLE OF CONTENTS

PREFACE

This article is a survey of an emerging subject, "Surfaces in Computer Aided Geometric Design". The purpose of this article is to present some of the fundamental concepts of our subject and to provide pointers to other work, enabling the reader to pursue the subject further. (In our subject many of the results are very new and others are "folklore", making scholarly study difficult.)

INTRODUCTION AND HISTORY

Significance of Surfaces in CAGD

Most scientific representation of information requires approximations at some level. The approximation might occur at the level of equations that model the physical reality, or at the level of the numerical solution of these equations.

As in any science, for creating surfaces one has some quantitative data (such as scientific measurements) and some qualititative information (such as intuition of a "good" shape). The quantitative data can be thought of as "hard" data such as given positions and tangents. The qualitative data may be thought of as "soft" information such as the desired shape. The philosophy for the construction of surfaces can be either interpolation or approximation. Interpolation means that one matches the given data exactly and approximation, a more general term, means one nearly matches the data. This dichotomy is discussed at some length in P. J. Davis' (1975) book.

Interpolation and Approximation.

At the most general level the tools employed to create surfaces include differential geometry, numerical analysis and computer graphics. Differential geometry is used to define surfaces. The spirit of numerical analysis is used to define surface interpolation methods to

display surface forms efficiently by means of computer graphics (Barnhill, 1983b). Computer graphics itself is an important research area which has undergone much growth in the past few years (Newman and Sproull, 1979; Foley and Van Dam, 1982). Computer graphics illustrations play a central role in understanding and evaluating surfaces. A graphical capability that is tailored to surface schemes makes possible an immediate presentation of results with minimal interaction by the user. This wedding of mathematics and technology makes the subject more useful and more difficult.

History: Surfaces in Computer Aided Geometric Design

The representation and approximation of surfaces in a computer graphics environment may be considered to have been launched by two pioneers: S. A. Coons and P. Bezier. Coons'(1964) surfaces and Bezier's[1] (1966, 1967) surfaces each consist of a network of "patches" which have a rectilinear topology. Coons' patches match exactly certain information (namely, whole curves of data). Bezier's surface methods have the different flavor that some data are matched exactly and the rest are approximated. Thus Coons' patches are a form of interpolation and Bezier's patches are a form of approximation which corresponds at a high level to Davis' dichotomy of interpolation and approximation. Specifically, Coons' "blending functions" are the basis functions for Hermite interpolation and Bezier's blending functions are the basis functions for Bernstein approximation. Barnhill (1982) briefly surveys Coons patches and Barnhill (1985) and Farin (1985b) are preparing extensive surveys of Coons and Bezier patches, respectively.

Both Coons and Bezier were working in engineering environments when they discovered their patch methods. In order for mathematicians to analyze their methods, the underlying structures of the methods had to be

recognized. As we shall see, their basic methods have been generalized and improved in various ways.

W. J. Gordon (1969a) discovered that Coons' patches have the powerful underlying algebraic structure of forming distributive lattices. Gordon described Coons' patches as Boolean sums of lower-dimensional "projectors" which were themselves interpolants to lower-dimensional information.

At about the same time "Gordon surfaces" (Gordon, 1969b, 1969c) consisting of a network of patches were created. Gordon surfaces blend together a given rectilinear network of curves. The blending can be achieved, for example, with univariate Lagrange interpolants or interpolatory splines.

A generalization of Coons' original patches was also necessary for those situations in which four-sided topology cannot be assumed. Barnhill, Birkhoff and Gordon (1973) initiated triangular Coons' patches for the case of arbitrarily located information. This innovation created many new surface possibilities. These triangular methods have a more complicated data structure through which they solve the more complex problem of interpolation to more general data. Subsequently additional triangular patches have been discovered (Barnhill, 1983a, 1983b; Nielson and Franke, 1983).

There has been a parallel set of developments for Bezier's methods: 1) Gordon showed how Coons' patches could be analyzed mathematically. The corresponding discovery for Bezier's patches was done by Forrest (1972) who showed that Bezier curves and surfaces could be considered as Bernstein polynomial approximations. This recognition has made possible the discovery of many important features of Bezier appoximations, such as the convex hull property and the variation diminishing property.

2) The analogue to Gordon surfaces is a network of rectangular Bezier patches: tensor product B-splines were discussed by Gordon and Riesenfeld (1974).

3) The analogue to the Barnhill, Birkhoff, and Gordon triangular Coons patch is the triangular Bezier patch which, for an arbitrary triangle, was discovered by Farin (1980)[2]. Farin's generalization has opened up many possibilities for the creation of new triangular interpolants as well as useful descriptions of known triangular interpolants. In fact, the Bezier method has become the starting point for generalizations that develop piecewise polynomial schemes with desired geometric properties, as is mentioned by several authors in this Volume.

The history of Coons patches and Bezier patches is summarized in Figure 1. (The idea of this Figure was conceived jointly with G. Farin who also made the drawing.)

Place Figure 1 here.

Caption for Figure 1: The history of Coons patches and Bezier patches.

Computer Aided Geometric Design

The term "Computer Aided Geometric Design" was invented by R. E. Barnhill and R. F. Riesenfeld in 1974 to describe the mathematical aspects of Computer Aided Design (hence the word "geometric"). The term first appeared as the title of the symposium held at The University of Utah and the subsequent book published by Academic Press. Computer Aided Geometric Design focuses on design. In order to recognize the need for a new emphasis on representation and to focus on surfaces instead of curves, the new term "Surfaces in Computer Aided Geometric Design" was coined by

Barnhill, Boehm and Farin in 1981.

Surfaces in Computer Aided Geometric Design

A number of additional significant changes, central to the direction of the field, are embodied in the new name, "Surfaces in Computer Aided Geometric Design." Let us make these explicit here.

1. The research focuses on surfaces, not on curves.
2. The surfaces, moreover, need not be built up from curves.
3. Geometric data for surfaces can be arbitrarily located. (Surfaces used in practice have usually been based on rectangularly structured tensor product data.)
4. Multidimensional surfaces are investigated.

ENVIRONMENT

Scholarly Environment for a New Subject

Disciplines that have strong technological components tend to be pursued in fragmentary ways with each problem treated on an ad hoc basis. "Surfaces in CAGD" is an example of such a discipline. The subject can be made more scientific and integrated by means of research, training of new professionals in the field (Keyworth, 1983), collaborations, research symposia such as this one, books, and journals. Several books have summarized the research in this area: Computer Aided Geometric Design, edited by Barnhill and Riesenfeld (1974), Surfaces in Computer Aided Geometric Design, edited by Barnhill and Boehm (1983) and the Surfaces issue of the Rocky Mountain Journal of Mathematics, edited by Barnhill and Nielson (1984). The new journal, Computer Aided Geometric Design, is devoted to recent research in this area.

SURFACES

Choice of Surface Form: Applications

Surfaces in Computer Aided Geometric Design have many applications, including fitting experimental data, tables of numbers and discretized solutions of differential equations; the design of aircraft, cars, and many other objects; and modeling human organs and robots. The term "Computer Aided Design/Computer Aided Manufacturing" (CAD/CAM) is used to describe some of these applications, particularly in engineering. The choice of the surface form depends upon the application, that is, there is no single solution for all problems. The variety of applications is so great that there cannot be a universal panacea. For example, the surface form used to model the human heart is unlikely to have the correct properties for modeling a car body. Consequently, we shall consider several families of methods in both interpolation and approximation senses. We use the term "surface modeling" to describe all applications since in all cases the mathematics describes a physical model.

Design and Representation of Surfaces

"Surfaces in CAGD" has two main categories: the Design of Surfaces and the Representation of Surfaces. Design of Surfaces involves making interactive changes in surfaces and displaying the surface in real-time. Representation of Surfaces involves using information derived elsewhere and of viewing the surface in order to understand its properties. These two categories have some common and some different features.

The features common to both Design and Representation include:

1) Some smoothness is desirable. (This smoothness might be C^0, C^1, or C^2 continuity or might be "visual continuity" of some order.)

2) Shape fidelity must be satisfied. (This may be somewhat vague,

such as a designer's idea of "sweetness" or a geophysicist's view of what a surface should look like.)

3) Methods may be either local or global. (With local methods the evaluation of the surface depends only on nearby data.)

4) Computer graphics are useful.

Features that differ for Design and Representation include:

1) The data can be modified and, possibly, augmented for Design. The data for Representation usually are fixed and are expensive to obtain, e.g., results from wind tunnel tests. (However, the representation used can affect the design of the experiments.)

2) Design surfaces are usually in three-space ("three-dimensional surfaces"), but Representation can take place in n-space.

3) Finally, Design involves computer graphics allowing real-time modification so that the designer can get immediate feedback, whereas Representation involves viewing surfaces in order to understand them but not necessarily to make interactive changes in them. These functions may be seen as editing surfaces and viewing surfaces, respectively.

Coons patches and Bezier patches

We now discuss Coons patches and Bezier patches. As mentioned above, patches can be either rectangular or triangular. Let us begin with rectangular Coons patches. The first case is the bilinearly blended Coons patch which interpolates to four curves as shown in Figure 2.

Place Figure 2 here.

Caption for Figure 2:

Four data curves for the bilinearly blended Coons patch.

Gordon pointed out that this patch can be written as a Boolean sum of linearly ruled lofting interpolants, more precisely, if

$$P_1 = P_1F = (1-u)\ F(0,v) + u\ F(1,v)$$

and P_2 is defined analogously, then the Boolean sum defined by

$$P_1 \oplus P_2 = P_1 + P_2 - P_1\ P_2$$

interpolates to the four boundary curves. We observe the nice interplay between algebra and geometry here: one can build up the idea of the Boolean sum by looking at P_1, P_2, and $P_1\ P_2$, thus using geometry, and then verify the correctness of the solution by direct substitution, thus using algebra.

Next let us consider the bicubically blended Coons patch which can be built up algebraically in an analogous way, namely, by means of the Boolean sum of univariate projectors. The projector P_1 is given by $P_1F =$

$$h_0(u)\ F(0,v) + h_1(u)\ F(1,v) + \bar{h}_0(u)\ F_{1,0}(0,v) + \bar{h}_1(u)\ F_{1,0}(1,v)$$

where the blending functions h_0, h_1, $\bar{h}_0$, and $\bar{h}_1$ are the cubic Hermite basis functions and $F_{1,0}$ means partial derivative with respect to u. The projector P_2F is defined similarly in the second variable v.

The term $P_1\ P_2F$ involves the following data:

positions	tangents
tangents	twists

The twists, which are (1,1)-derivatives, can cause problems in using this patch. Solutions to this problem are given by Gregory (1974) and by Barnhill, Brown, and Klucewicz (1978). An application of their research to reducing surface oscillations by varying twists is given by Brunet in this Volume.

We recently created the multidimensional compatibly corrected C^1 Coons

patch for a higher dimensional representation problem (Barnhill and Worsey, 1984). The generalization to C^2 Coons patches is discussed by Worsey in this Volume.

Coons patches are "transfinite"; this term, introduced by Gordon (1971), connotes that whole curves of data are interpolated. These data can be discretized leading to finite dimensional interpolants some of which are called "serendipity elements" in the finite element literature. An example is the discretization of the bicubically blended Coons patch to the standard 16 degree of freedom bicubic patch obtained by replacing $F(u,0)$, $F_{0,1}(u,0)$, etc. by their respective cubic Hermite interpolants.

Triangular Coons Patches

Triangular Coons patches were initiated by Barnhill, Birkhoff, and Gordon (1973) who considered the C^1 case with the corresponding cubic Hermite projectors along parallels to each side, that is, P_1 is the cubic Hermite lofting interpolant along parallels to side 1 etc. The "BBG triangle" is a family of interpolants formed by taking Boolean sums of the three lofting interpolants P_1, P_2, and P_3. Twist incompatibilities inherent in all Boolean sums must again be resolved in order to produce suitable schemes. Little (1978) made the very important step of generalizing the "BBG" schemes to an arbitrary triangle (Barnhill and Little, 1984), a key concept being a calculus for functions of barycentric coordinates.

Other transfinite triangular interpolants have subsequently been discovered, including Nielson's radial schemes (Nielson, 1979), Gregory's symmetric schemes (which are generalized to n-dimensional simplices in this Volume), and Brown, Dube, and Little's convex combinations. Recently Alfeld and Barnhill (1984) constructed a C^2 BBG scheme. Finally, Little has

devised a trivariate C^1 BBG scheme (Barnhill and Little, 1984).

Bezier Patches

Bernstein-Bezier approximations have recently become very popular and no fewer than 1/4 of the titles at this symposium contain Bezier's name. Bernstein-Bezier patches interpolate some data and approximate others. This representation is described by a "net" of "control vertices", where the vertices of the net are the coefficients of the Bernstein basis functions. (See Figure 1 in Farin's article in this Volume.)

Bezier patches (like Coons patches) can be either rectangular or triangular. Rectangular Bezier patches are tensor products, so their properties follow from the univariate case. (For additional information on tensor products, as well as on CAGD in general, see Boehm, Farin, and Kahmann, 1984.) Therefore, we shall content ourselves with a brief introduction to triangular Bezier patches.

A necessary tool for the construction of a triangular interpolant over an arbitrary triangle is the concept of barycentric coordinates. The barycentric coordinates of the arbitrary point P in a triangle with vertices 1, 2, 3 are given by $b_i = A_i/A$ where A is the area of the triangle and A_i is the area of the subtriangle opposite vertex i, i = 1,2,3. This geometry for barycentric coordinates appears in Figure 3. The definition of barycentric coordinates implies that they are non-negative if P is in the triangle and that they sum to one.

Place Figure 3 here.

Caption for Figure 3:

Geometry for the barycentric coordinates of the point P.

The triangular Bernstein polynomial is given by

$$\sum_{i+j+k=n} \frac{n!}{i!j!k!} \quad b_1^i \, b_2^j \, b_3^k \, V_{i,j,k}$$

where the $V_{i,j,k}$ are (vector-valued) "control vertices." Imposing continuity between Bezier patches is important. This continuity has striking geometric interpretations, for example, to obtain C^1 continuity certain quadrilaterals must be planar. See Figures 2a and 2b in Farin's article in this Volume for the C^1 and C^2 cases, respectively.

Interpolants for CAGD are most convenient if written in a form in which the data occur explicitly, i.e., "cardinal" form. The Farin generalization of Bernstein-Bezier approximation is particularly useful for finding the cardinal form of interpolants, both old and new. We mention several useful polynomial triangular interpolants whose cardinal forms have been found explicitly using Farin's methods: the C^1 "Clough-Tocher triangle" (Clough and Tocher, 1965; Farin, 1980), the C^1 quintic "macro-triangle" with C^2 data at its vertices (Strang and Fix, 1973; Barnhill and Farin, 1981), a C^2 Clough-Tocher triangle discovered by Alfeld (1984c) and a C^2 macro-triangle (Zenisek, 1970, 1973; Kolar et al., 1971) given explicitly by Whelan (1985) and by Rescorla (1985). We expect additional interpolants to be discovered using Farin's methods. We think that the reason these methods enable interpolants to be more easily analyzed than earlier methods is that the effects of specifying the geometry or the algebra are readily apparent, e.g., the first row in a Bezier net concerns continuity, the second row C^1 continuity, etc. For a comparison between Bezier and Coons methods for finding one interpolant see Barnhill and Farin (1981).

In order to apply smooth triangular interpolants for interpolation to arbitrarily located data two preprocessing steps are needed: 1) triangulation of the domain data sites and 2) estimation of derivatives needed by the interpolants.

Triangulation

Triangulations of a given set of points are not unique. Moreover, several triangulation methods are known. We recommend the following algorithm which is due to Little and to Petersen:

1) Specify the boundary, the default being the convex hull.
2) Create an initial triangulation.
3) Optimize this triangulation using Lawson's exchange algorithm with the criterion min max angle where the minimum is over all triangulations and the maximum over all angles in a triangulation T, that is, $\underset{\text{all triangulations } T}{\text{minimize}} \quad \text{maximum angle.}$

For additional information on this algorithm, see Barnhill (1977), Lawson (1977), Barnhill (1983a), and Barnhill and Little (1984).

An example is shown in Figures 4a and 4b. A specified (curved) boundary, including holes, and specified data sites to be triangulated are given in Figure 4a. The triangulation resulting from the Little-Petersen algorithm is displayed in Figure 4b.

Place Figures 4a and 4b here.

Caption for Figures 4a and 4b:

Triangulation: The boundary and data sites are specified.

Then the data sites are triangulated with the boundary maintained.

Estimation of Derivative Data

Users ordinarily supply only positional information. Most surface methods require some derivative data which must usually be created. A survey of the topic of derivative estimation is given in Stead (1984). Three useful possibilities for the estimation of derivatve data are the following:

1) "Triangular Shepard's Method" (Little, 1983; Piper, 1983) is of the form $\sum_i w_i L_i F$ where $w_i = A_i / \sum_j A_j$, $A_i = 1/[d_{i1} d_{i2} d_{i3}]^2$, $L_i F$ is the linear interpolant over the ith triangle and d_{ik} is the distance from a fixed point to the kth vertex of the ith triangle, k = 1,2,3.
   (The original "Shepard's Method" is given in Shepard, 1968.)
2) "Hardy's Multiquadrics" (Hardy, 1971; Franke, 1982) are a family of schemes one example being $\sum_i c_i [d_i^2 + R]^{1/2}$.
3) The coefficients of an interpolant are computed by minimizing a (nonlinear) functional such as the L_2 norm of the second derivatives (Alfeld, 1983).

A note on splines and norms: Splines can be constructed as the solution of variational problems, more precisely, as interpolants of minimum norm in the relevant function spaces (de Boor, 1978). The Sobolev seminorm involving second derivatives, that is, the norm used by Alfeld above, is also used to find the "thin plate splines" discussed by Franke (1985) in this Volume. A similar but different norm for surface approximation is used for "surfaces under tension" (Nielson and Franke, 1984). Additional examples are the rectangular nu-splines of Nielson and the tau-splines of Hagen in this Volume. (Yet another norm is used in Nielson, 1980.)

Contouring

Contouring, rather than finding a surface per se, sometimes is the actual problem to be solved. (Recall that a contour of, say, $w = F(x,y,z)$ is the set of (x,y,z) whose image is a given w.) Examples of contouring include maps, silhouette edges, and hidden surfaces. Little (1981) created an adaptive subdivision and degree lowering algorithm which Petersen (1983, 1984) implemented. Intrinsic use is again made of Farin's representation of polynomials over an arbitrary triangle, in particular, the convex hull property of this representation. Little suggests starting with a cubic polynomial over a triangle and reducing its degree to linear polynomials over subtriangles, since linear functions are trivial to contour. Petersen's adaptation of the algorithm to surfaces in 4-space is used to generate the three-dimensional contours shown in Figure 5. (A different application of this general degree raising and lowering algorithm, namely, to finding hidden surfaces, is made by Arner, 1985.)

The two test functions in Figure 5 are the monomial $F(x,y,z) = xyz$ and the "trigonometric function" $g(x,y,z) = \cos(3.14x)\cos(y-0.5)\sin[3.14(z-0.5)]$. Related, different color pictures are given in Barnhill and Stead (1984).

Place Figure 5 here.

Caption for Figure 5:

Contours of two four-dimensional surfaces. The contours are coded by color. Notice that a contour can have several branches.

Four-Dimensional Surfaces

We have recently been asked to consider "surfaces" in higher-dimensional spaces, for example, the data may be of the form $(x_i, y_i, z_i; T_i)$ where x,y,z are the usual 3-space coordinates and T represents the corresponding

temperature, that is, the functional relationship is assumed to be of the form $T = T(x,y,z)$. The function T is a trivariate function and the corresponding surface in 4-space is called a "four-dimensional surface" for simplicity. Problems that may be usefully interpreted as higher-dimensional surfaces occur often in applications, although the user may not be aware that "surfaces" are involved. Higher-dimensional surfaces that we have encountered include the above temperature problem as well as some five- and six-dimensional combustion problems with Rosemary Chang at Sandia Laboratories, trivariate metallurgical problems of alloy-mixing with Don Orser at the National Bureau of Standards, and the oblique wing problem with Sarah Stead at NASA-Ames. The rationale for multivariate approximations, as well as additional examples with particular emphasis on Department of Energy applications, are given by Barnhill and Chang (1979).

We focus on four-dimensional surfaces here. In order to construct four-dimensional surfaces, we suggest the following high-level algorithm (Barnhill and Little, 1984):

1) Triangulate the domain data into optimal tetrahedra.
2) Create necessary derivative information at the data sites.
3) Interpolate smoothly to the dependent variable (e.g., the T_i) using the tetrahedral interpolants of Barnhill and Little (1984), Alfeld (1984a, 1984b), Farin (1985b), or Gregory in this Volume.
4) Render the resulting four-dimensional surface, perhaps using its (three-dimensional) contours. We index these contours by color, as illustrated in Figure 4. Additional examples of this use of color are in Barnhill and Stead (1984).

In one special form of this general problem the domain points lie on a (known) three-dimensional surface. We recently considered an aerodynamic

problem with Stead at NASA-Ames (Barnhill, Piper, and Stead, 1984, 1985) that exemplifies this variation. In this application the problem is to represent the pressure at any point on an oblique wing, given the pressures at points with a certain "structure." Thus the pressure "surface" is defined on the surface of the wing. Additional information on this problem is given in the paper by Barnhill, Piper, and Stead in this Volume.

Multistage Methods

"Multistage methods" utilize the successive application of several surface schemes to solve a given problem. We first read this idea in Schumaker (1976). It has been developed by Foley and Nielson (1980), Stead (1983), and Barnhill and Stead (1984). An elementary example of a multistage method is the composition HS where S is a Shepard's method (Barnhill, Dube, and Little, 1983) and H is tensor product Hermite interpolation. This example illustrates why multistage methods can be useful:

1) S interpolates the given arbitrarily located data, but S is global which implies that its value at a point can be affected by data far away. Localizing a global interpolant directly is expensive, so S itself is unsuitable.

2) Given the surface corresponding to S, the tensor product H can be applied to S. We note that H could not be applied to the original arbitrarily located data because H requires rectilinear data as input. However, H has the advantage of being a local approximation. A reason for using a tensor product approximation such as H is that rendering a surface requires many evaluations and, moreoever, the evaluation sites are frequently on a rectilinear grid, e.g., many plotting routines require evaluations on a rectilinear grid.

Boolean sums are multistage methods which can be used to achieve

interpolation and polynomial precision according to the Barnhill/Gregory Theorems (Barnhill and Gregory, 1975). These theorems have been used to construct interpolants with desired properties by Barnhill and Gregory (1975) and by Poeppelmeier (1975) with further applications given in Barnhill, Dube, and Little (1983) and to Foley and Nielson's (1980) "delta sum iteration."

Future Research: Open Research Questions

We conclude our survey by presenting some open research questions that appeal to us. This list does not exhaust our field's rich possibilities for future research.

1) Rendering of 4-dimensional Surfaces.

   Many problems are equivalent to finding four-dimensional (or higher-dimensional) surfaces. We render our resulting four-dimensional surfaces by means of color-coded (three-dimensional) contours. Alternative possibilities for representing the fourth dimension in a wide variety of problems are needed.

2) Tetrahedral Interpolants.

   Smooth interpolants defined over tetrahedra are currently being developed. BBG transfinite interpolants with their discretizations and n-dimensional radial Nielson interpolants are given in Barnhill and Little (1984). A transfinite interpolant formed by means of a convex combination of BBG projectors is presented by Alfeld (1984a). A C^1 Clough-Tocher tetrahedral interpolant was discovered by Alfeld (1984b) and a second by Farin (1985b). Gregory's triangular "symmetric scheme" (another convex combination of

projectors) has been generalized to n-dimensional simplices by Gregory (1979, 1985). (See also Mansfield, 1976.)

These tetrahedral interpolants need evaluation and testing, part of the problem being to find suitable trivariate test examples.

3) Surfaces Defined on Surfaces.

The task is to use knowledge of the known domain surface to create suitable higher-dimensional surfaces defined on this domain. We cited our NASA oblique wing example above to illustrate this class of problems. A second interesting example is interpolation defined on a sphere (Lawson, 1984).

4) Optimal Triangulation of Tetrahedra.

There are two principal classes of methods for "triangulation" of points in 3-space into tetrahedra: 1) the min max or max min algorithms, one example being to maximize over all triangulations the minimum over all tetrahedra in a triangulation of R_I/R_C where R_I is the radius of the inscribed sphere and R_C is the radius of the circumscribed sphere (Petersen, 1983; Barnhill and Little, 1984) and 2) the Delaunay triangulation corresponding to the Dirichlet tessellation (Bowyer, 1981; Watson, 1981).[3] These two classes of methods have not been compared for tetrahedral triangulation. For planar triangulations they have different good and bad points: 1) the min max criteria are based on error bound analyses such as Barnhill and Gregory (1976a, 1976b) and Gregory (1975).[4] However, convergence of the algorithm is difficult to prove for some examples. 2) The Delaunay triangulation of points in the plane has been shown to be equivalent to (certain) max min criteria (Lawson, 1977; Sibson, 1978). Max min criteria are not

optimal according to the above error bounds (Caution: Max min criteria are optimal according to earlier, less sharp error bounds involving factors such as $1/\sin^n$ (min angle) where n-1 is the order of the method.) However, Delaunay algorithms always converge. The min max algorithms are, in fact, examples of Lawson exchange algorithms, which find local optima. Dirichlet tessellations produce global optima. These classes of methods must be studied further in order to determine when each is useful.

5) Monotone and Convex Surfaces.

Many physical phenomena involve surfaces that are monotone (in some directions) or are convex, an example being the modeling of equations of state. The corresponding representation problem is given some monotone or convex data, find a monotone or convex interpolant. Standard interpolants need not be monotone or convex even though they match monotone or convex data. Fritsch and Carlson (1983) have used bicubic interpolants defined over rectangles which are monotone for monotone data. Fritsch gave a Progress Report at this symposium. Gregory and Delbourgo (1982) have constructed monotone curves to monotone data which show promise of generalization to surfaces. Finally, convexity of the Bezier polygon can imply convexity of the Bezier approximation. This idea is investigated for triangular schemes by Chang and Davis (1984), Chang and Feng (1985), and Barnhill and Whelan (1985).

6) Visual Continuity.

Visual or geometric continuity refers to the perceived smoothness of a surface instead of to parametric continuity (which is an artifact of parametrization). Nielson (1974) initiated visual

continuity for planar curves with his polynomial alternatives to splines under tension which are parametrically C^1 but not C^2 but for which "d^2y/dx^2" is continuous. A generalization of Nielson's nu-splines to torsion continuity is given by Hagen in this Volume. The definition of second order visual continuity for space curves is currently under discussion. One possibility is curvature continuity, an example of which is Farin's (1982b) B-spline-like scheme. Another possibility is curvature vector continuity proposed by some of the students in the Math CAGD class at the University of Utah in 1984 and by Barsky and DeRose (1984). We think that these definitions do not cover all cases (Barnhill, Jordan, and Piper, 1984).

First order visual continuity for surfaces is tangent plane continuity and was considered by Coons (1964) and Bezier (1967, 1972). The definition of second order visual continuity for surfaces is still being developed. One possibility is continuity of the Dupin indicatrix (Kahmann, 1983).

7) Closed Surfaces.

General closed surfaces are not images of planar domains and so extra care is required in their construction. Herron (1979, 1984) and Farin (1983) have developed "domainless" triangular interpolants for modeling closed surfaces. Very little is known about the construction of general closed surfaces.

8) Rectangular and Triangular Patches.

Truly general design systems include both rectangular and triangular patches, but these different types of patches are difficult to blend together. "Domainless" interpolants such as

those developed by Gregory (1983), Gregory and Charrot (1980), and Farin (1982a) are useful in blending rectangular and triangular patches.

Acknowledgments

The research reported in this survey has been going on for about a decade now. Many people have contributed during this time and they all deserve thanks. This particular document has been improved by the editing of Marigold Linton, Rosemary Chang and Gerald Farin and by the (usual) vigorous discussions with the members of the Math CAGD Group at Utah. The research has been supported in part by the Department of Energy through contract DE-AC02-82ER12046 to The University of Utah.

REFERENCES

P. Alfeld (1983), Multivariate scattered data derivative generation by functional minimization (submitted for publication). Temporary reference: Mathematics Research Center Report 2703, University of Wisconsin, Madison, Wisconsin.

P. Alfeld (1984a), A discrete C^1 interpolant for tetrahedral data, Rocky Mountain Journal of Mathematics, 14, 5-16.

P. Alfeld (1984b), A trivariate Clough-Tocher scheme for tetrahedral data, Computer Aided Geometric Design, 1, 169-181.

P. Alfeld (1984c), A bivariate C^2 Clough-Tocher scheme, Computer Aided Geometric Design, 2 (to appear).

P. Alfeld and R. E. Barnhill (1984), A transfinite C^2 interpolant over triangles, Rocky Mountain Journal of Mathematics, 14, 17-40.

P. R. Arner (1985), Hidden surface elimination, Masters thesis, Department of Mathematics, University of Utah, Salt Lake City, Utah.

R. E. Barnhill (1977), Representation and approximation of surfaces, in: J. R. Rice, editor, Mathematical Software III, Academic Press, New York.

R. E. Barnhill (1982), Coons' patches, Computers in Industry, 3 (2), 37-43.

R. E. Barnhill (1983a), Computer aided surface representation and design, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

R. E. Barnhill (1983b), A survey of the representation and design of surfaces, IEEE Computer Graphics and Applications, 3 (7), 9-16.

R. E. Barnhill (1985), A survey of transfinite patch methods, in preparation.

R. E. Barnhill, G. Birkhoff, and W. J. Gordon (1973), Smooth interpolation in triangles, Journal of Approximation Theory, 8 (2), 114-128.

R. E. Barnhill and W. Boehm, editors (1983), Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

R. E. Barnhill, J. H. Brown, and I. M. Klucewicz (1978), A new twist for computer aided geometric design, Computer Graphics and Image Processing, 8, 78-91.

R. E. Barnhill and R. Chang (1979), Approximation, in: R. E. Huddleston, editor, Program Directions for Computational Mathematics, Department of Energy, Washington, D.C.

R. E. Barnhill, R. P. Dube, and F. F. Little (1983), Properties of Shepard's surfaces, Rocky Mountain Journal of Mathematics, 13 (2), 365-382.

R. E. Barnhill and G. Farin (1981), C^1 Quintic interpolation over triangles: two explicit representation, International Journal for Numerical Methods in Engineering, 17, 1763-1778.

R. E. Barnhill and J. A. Gregory (1975), Polynomial interpolation to boundary data on triangles, Mathematics of Computation, 29, 726-735.

R. E. Barnhill and J. A. Gregory (1976a), Sard kernel theorems on triangular domains with application to finite element error bounds, Numerische Mathematik, 25, 215-229.

R. E. Barnhill and J. A. Gregory (1976b), Interpolation remainder theory from Taylor expansions with nonrectangular domains of influence, Numerische Mathematik, 25, 401-408.

R. E. Barnhill, M. Jordan, and B. R. Piper (1984), Visual continuity for space curves and surfaces, Math CAGD Seminars, University of Utah, Salt Lake City, Utah.

R. E. Barnhill and F. F. Little (1984), Three- and four-dimensional surfaces, Rocky Mountain Journal of Mathematics, 14, 77-102.

R. E. Barnhill and G. M. Nielson, editors (1984), Surfaces, a special issue of the Rocky Mountain Journal of Mathematics, 14 (1).

R. E. Barnhill, B. R. Piper, and S. E. Stead (1984), Surface representation for the graphical display of structured data, Final Report on Joint Research Interchange with The University of Utah, NASA-Ames Research Center, Moffett Field, CA.

R. E. Barnhill, B. R. Piper, and S. E. Stead (1985), A multidimensional surface problem: pressure on a wing, this Volume.

R. E. Barnhill and R. F. Riesenfeld, editors (1974), Computer Aided Geometric Design, Academic Press, New York.

R. E. Barnhill and S. E. Stead (1984), Multistage trivariate surfaces, Rocky Mountain Journal of Mathematics, 14, 103-118.

R. E. Barnhill and T. Whelan (1985), A geometric interpretation of convexity conditions for surfaces, Computer Aided Geometric Design, 2 (to appear).

R. E. Barnhill and A. J. Worsey (1984), Smooth interpolation over hypercubes, Computer Aided Geometric Design, 1, 101-113.

B. A. Barsky and T. D. DeRose (1984), Geometric continuity of parametric curves, Report No. UCB/CSD 84/205, Computer Science Division (EECS), University of California, Berkeley, CA 94720.

P. Bezier (1966), Definition numerique des courbes et surfaces, Automatisme, 11, 625-632.

P. Bezier (1967), Definition numerique des courbes et surfaces (II), Automatisme, 12, 17-21.

P. Bezier (1972), Numerical Control, Mathematics and Application, Wiley, New York.

W. Boehm, G. Farin, and J. Kahmann (1984), A survey of curve and surface methods in CAGD, Computer Aided Geometric Design, 1, 1-60.

C. de Boor (1978), A Practical Guide to Splines, Springer-Verlag, New York.

A. Bowyer (1981), Computing Dirichlet tessellations, The Computer Journal, 24, 162-166.

P. Brunet (1985), Increasing the smoothness of bicubic spline surfaces, this Volume.

P. de Casteljau (1959), Outillage methodes calcul, Andre Citroen Automobiles SA, Paris.

P. de Casteljau (1963), Courbes et surfaces a poles, Andre Citroen Automobiles SA, Paris.

G. Chang and Y. Feng (1985), An improved condition for the convexity of Bernstein-Bezier surfaces over triangles, Computer Aided Geometric Design, 2 (to appear).

G. Chang and P. J. Davis (1984), The convexity of Bernstein polynomials over triangles, Journal of Approximation Theory, 40, 11-28.

R. W. Clough and J. L. Tocher (1965), Finite element stiffness matrices for analysis of plates in bending, Proceedings of Conference on Matrix Methods in Structural Mechanics, Air Force Institute of Technology, Wright-Patterson A.F.B., Ohio.

S. A. Coons (1964), Surfaces for computer aided design, Mechanical Engineering Department, M.I.T., revised, 1967. (Available as AD 663 504 from the National Technical Information Service, Springfield, VA 22161.)

P. J. Davis (1975), Interpolation and Approximation, Dover, New York.

G. Farin (1980), Bezier polynomials over triangles and the construction of piecewise C^1 polynomials, Department of Mathematics TR/91, Brunel University, Uxbridge England.

G. Farin (1982a), A construction for the visual C^1 continuity of polynomial surface patches, Computer Graphics and Image Processing, 20, 272-282.

G. Farin (1982b), Visually C^2 cubic splines, Computer Aided Design, 14, 137-139.

G. Farin (1983), Smooth interpolation to scattered 3D data, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

G. Farin (1985a), A modified Clough-Tocher interpolant, this Volume.

G. Farin (1985b), A survey of Bernstein-Bezier triangular interpolants, in preparation.

I. D. Faux and M. J. Pratt (1979), Computational Geometry for Design and Manufacture, Halsted Press, New York.

T. A. Foley and G. M. Nielson (1980), Multivariate interpolation to scattered data using delta iteration, in: E. W. Cheney, editor, Approximation Theory III, Academic Press, New York.

J. D. Foley and A. Van Dam (1982), Fundamentals of Interactive Computer Graphics, Addison-Wesley, Reading, MA.

A. R. Forrest (1972), Interactive interpolation and approximation by Bezier polynomials, The Computer Journal, 15, 71-79.

R. H. Franke (1982), Scattered data interpolation: test of some methods, Mathematics of Computation, 38, 181-200.

R. H. Franke (1985), Thin plate splines with tension, this Volume.

F. N. Fritsch (1985), Monotonicity preserving bicubic interpolation, this Volume.

F. N. Fritsch and R. E. Carlson (1983), Monotone piecewise bicubic interpolation, SIAM Journal of Numerical Analysis (submitted). Temporary reference: Lawrence Livermore Laboratory Preprint 86449, Rev. 1.

W. J. Gordon (1969a), Distributive lattices and the approximation of multivariate functions, in: I. J. Schoenberg, editor, Approximations with Special Emphasis on Splines, University of Wisconsin Press, Madison.

W. J. Gordon (1969b), Free-form surface interpolation through curve networks, Research Publication GMR-921, General Motor Research Labs., Warren, MI 48090.

W. J. Gordon (1969c), Spline-blended surface interpolation through curve networks. Journal of Mathematics and Mechanics, 18, 931-952.

W. J. Gordon (1971), Blending-function methods of bivariate and multivariate interpolation and approximation, SIAM Journal Numerical Analysis, 8, 158-177.

W. J. Gordon and R. F. Riesenfeld (1974), B-spline curves and surfaces, in: R. E. Barnhill and R. F. Riesenfeld, editors, Computer Aided Geometric Design, Academic Press, New York.

P. J. Green and R. Sibson (1978), Computing Dirichlet tessellations in the plane, The Computer Journal, 21, 168-173.

J. A. Gregory (1974), Smooth interpolation without twist constraints, in: R. E. Barnhill and R. F. Riesenfeld, editors, Computer Aided Geometric Design, Academic Press, New York.

J. A. Gregory (1975), Error bounds for linear interpolation on triangles, in: J. R. Whiteman, editor, The Mathematics of Finite Elements and Applications II, Academic Press, London.

J. A. Gregory (1979), Interpolation to boundary data on simplices, Department of Mathematics TR/87, Brunel University, Uxbridge, England.

J. A. Gregory (1983), C^1 rectangular and non-rectangular surface patches, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

J. A. Gregory (1985), Interpolation to boundary data on the simplex, this Volume.

J.A. Gregory and P. Charrot (1980), A C^1 triangular interpolation patch for computer aided geometric design, Computer Graphics and Image Processing, 13, 80-87.

J. A. Gregory and R. Delbourgo (1982), Piecewise rational quadratic interpolation to monotonic data, Institute of Mathematics and its Applications Journal of Numerical Analysis, 2, 123-130.

H. Hagen (1985), Geometric spline curves, this Volume.

R. L. Hardy (1971), Multiquadric equations of topography and other irregular surfaces, Journal of Geophysical Research, 76, 1905-1915.

G. J. Herron (1979), Triangular and Multisided Patch Schemes, Ph.D. thesis, Department of Mathematics, University of Utah, Salt Lake City, Utah.

G. J. Herron (1984), Smooth closed surfaces with discrete triangular interpolants, Computer Aided Geometric Design (under revision).

J. Kahmann (1983), Continuity of curvature between adjacent Bezier patches, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

G. A. Keyworth, II (1983), Federal R & D and industrial policy, Science, 220, 1122-1125.

V. Kolar, J. Kratochvil, A. Zenisek, and M. Zlamal (1971), Technical, Physical, and Mathematical Principles of the Finite Element Method, Academia, Czechoslovak Academy of Sciences, Prague, Czechoslovakia.

C. L. Lawson (1977), Software for C^1 surface interpolation, in: J.R. Rice, editor, Mathematical Software III, Academic Press, New York.

C. L. Lawson (1984), C^1 surface interpolation for scattered data on a sphere, Rocky Mountain Journal of Mathematics, 14, 177-202.

F. F. Little (1978), Private communication.

F. F. Little (1981), Tessellation of tetrahedra, Math CAGD Seminars, University of Utah, Salt Lake City, Utah.

F. F. Little (1983), Convex combination surfaces, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

L. Mansfield (1976), Interpolation to boundary data in tetrahedra with applications to compatible finite elements, Journal of Mathematical Analysis and Applications, 56, 137-164.

W. M. Newman and R. F. Sproull (1979), Principles of Interactive Computer Graphics, Second Edition, McGraw-Hill, New York.

G. M. Nielson (1974), Some piecewise polynomial alternatives to splines under tension, in: R. E. Barnhill and R. F. Riesenfeld, editors, Computer Aided Geometric Design, Academic Press, New York.

G. M. Nielson (1979), The side-vertex method for interpolation in triangles, Journal of Approximation Theory, 25, 318-336.

G. M. Nielson (1980), Minimum norm interpolation in triangles, SIAM Journal Numerical Analysis, 17, 44-62.

G. M. Nielson (1985), A rectangular nu-spline for interactive surface design, this Volume.

G. M. Nielson and R. H. Franke (1983), Surface construction based upon triangulations, in: R. E. Barnhill and W. Boehm, editors, Surfaces in Computer Aided Geometric Design, North-Holland, Amsterdam.

G. M. Nielson and R. H. Franke (1984), A method for construction of surfaces under tension, Rocky Mountain Journal of Mathematics, 14, 202-222.

C. S. Petersen (1983), Contours of three and four dimensional surfaces, Masters thesis, Department of Mathematics, University of Utah, Salt Lake City, Utah.

C. S. Petersen (1984), Adaptive contouring of three-dimensional surfaces, Computer Aided Geometric Design, 1, 61-74.

B. Piper (1983), Triangular Shepard's methods, Math CAGD Seminars, University of Utah, Salt Lake City, Utah.

C. C. Poeppelmeier (1975), A Boolean sum interpolation scheme to random data for computer aided geometric design, Masters thesis, Computer Science Department, University of Utah, Salt Lake City, Utah.

K. L. Rescorla (1985), Hermite interpolation over simplices, submitted for publication.

M. A. Sabin (1976), The use of piecewise forms for the numerical representation of shape, Tanulmanyok 60/1977, Hungarian Academy of Sciences, Budapest, Hungary.

A. Sard (1963), Linear Approximation, Mathematics Surveys Number 9, American Mathematical Society, Providence, R.I.

L. L. Schumaker (1976), Fitting surfaces to scattered data, in: G. G. Lorentz, C. K. Chui, and L. L. Schumaker, editors, Approximation Theory II, Academic Press, New York.

D. Shepard (1968), A two-dimensional interpolation function for irregularly-spaced data, Proceedings of the Association for Computing Machinery National Conference, 517-524.

R. Sibson (1978), Locally equiangular triangulations, The Computer Journal, 21, 243-245.

S. E. Stead (1983), Smooth multistage multivariate approximation, Ph.D. thesis, Division of Applied Mathematics, Brown University, Providence, R. I. 02912.

S. E. Stead (1984), Estimation of gradients from scattered data, Rocky Mountain Journal of Mathematics, 14, 265-280.

G. Strang and G. J. Fix (1973), An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, N. J.

D. F. Watson (1981), Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes, The Computer Journal, 24, 167-172.

T. Whelan (1985), A representation of a C^2 interpolant over triangles, submitted for publication.

A. J. Worsey (1985), C^2 interpolation over hypercubes, this Volume.

A. Zenisek (1970), Interpolation polynomials on the triangle, Numerische Mathematik, 15, 283-296.

A. Zenisek (1973), Polynomial approximation on tetrahedrons in the finite element method, Journal of Approximation Theory, 7, 334-351.

Footnotes

[1] Bezier and de Casteljau (1959, 1963) independently developed equivalent curve and surface schemes now known only under Bezier's name.

[2] Bezier patches over equilateral triangulations are given by Sabin (1976).

[3] An algorithm for Delaunay triangulation in the plane is given by Green and Sibson (1978).

[4] These error analyses are built on the work of Sard (1963).

These footnotes correspond to the following pages in the manuscript:

[1] page 2

[2] page 4

[3] page 18

[4] page 18

To appear in a book Ed by Roy Lasek

3.8.1

Draft

PARAMETERS OF DENDRITIC SHAPE AND SUBSTRUCTURE:
INTRINSIC OR EXTRINSIC DETERMINATION OF NEURONAL SHAPE?

Dean E. Hillman
Professor of Physiology and Biophysics
Department of Physiology and Biophysic
New York University Medical Center
550 1st Avenue, New York, NY
10016

Send Proofs to:

Dean E. Hillman
Department of Physiology and Biophysic
New York University Medical Center
550 1st Avenue, New York, NY
10016

Dendritic Shape and Substructural Determinants.

ABSTRACT

This review summarizes concepts of neuronal form as based on understandings of global parameters of neuronal dendritic arbors and their emergence from segmental parameters. Furthermore, evidence is provided for determinants of dendritic arbor parameters suggesting the sources of intrinsic controls defined by genetic expression of individual neurons and by extrinsic controls subtended by determinants of interactions between neurons.

A simplified model of arbors is used to define the segmental parameters that give ries to the global arbor shape. this model consists of two factors: 1) the dividing up of a total cross sectional area for dendritic processes that can be generated by a soma. Secondly this cross sectional area bifurcates until a limiting terminal cross sectional area is reached. Between the soma and the terminals segments of variable lengths distribute the cross sectional area between the bifuractions. The cross sectional area for each arbor can be modified between the soma and terminals as changes in the relative amount between the daughter branches or can be decreased or increased as taper along segments or at bifuractions (branch power). segments can be ordered according to the number of terminals that each segment has distal and is useful to make analysis of segment sizes, defining branch pattern classes, and in naming arbors for reconstruction. This model generates three categories of

segmental parameters, size (cross sectional area and length), orientation, and pattern of segments which in turn define the emergence of global parameters for arbors. Global parameters being much more difficult to define are best described in terms of parameters emerging from segments.

The determinants of shape parameters are from a complex interaction of intrinsically and extrinsically defined genetic expressions and from epigenetic influences that can alter the course of genetic expressions. Intrinsically, there appear to be four parameters that are determined intrinsically by the genetic expression from within an individual neuron. These are the Total amount of dendritic cross sectional area that a neuron can generate, the limit on terminal segment final bifuractions size, tapering from the soma to the terminals as segment taper as branch power and total volume that can be generated by the arbor skeleton. Numerous other parameters are determined by genetic expressions being subserved between neurons as interactions between specific neuronal types as well as with glia. These parameters are distribution of total cross sectional area for dendritic processes into stem segments of arbors, distribution of segment cross sectional area between daughter segments, length of segments, orientation of segments and pattern of segment interconnections.

Globally, these interactive parameters generate much of the neuronal shape seen in dendritic arbors and are guided by underlying constraits determined intrinsically by each individual neuron. These are due to specific subcellular structures.

3.8.4

The subcellular elements that are responsible for the intrinsically determined parameters of dendrites are neurotubules and neurofilaments. Neurotubules generate the skeleton on which other subcellular elements are supported. The number of neurotubules that can be generated by a soma represents the basis of the total cross sectional area of dendrites that can emerge from a soma. The smallest terminal segments that can be supported appear to be the result of a minimum number of neurotubules. The total volume of the arbor seems to be represented by the total amount of assembled tubulin generating length and cross sectional area to generate volume. Finally, the negative taper produced along arbors is related in part to the occurrence of neurofilaments.

1.0 INTRODUCTION

Neuronal form is determined from sources both intrinsic and extrinsic to each neuron (Rakic 1974). Those determinants, having an intrinsic influence, stem from within individual neurons and are primarily genetic in origin (Fig.1). This intrinsic driving force acts specifically through subcellular structures directing form internally through chemical and physical factors of development while outwardly specifying interactions for extrinsic control.

Extrinsic factors have both genetic and epigenetic sources. The extrinsic-genetic source arises by genetic expression from neurons and glia and directs interactions between individual neurons and also between neurons glia (Fig.1). This extrinsic-genetically determined order augments intrinsic expression of neuronal form through molecular interactions from two sources. These molecular interactions affect growth generating neuronal arbors and forming neuronal circuitry. This expression also provides maintenance connectivity well as directs plasticity due to functional and perturbative influences on the system.

The epigenetic influence, on the other hand, arises from environmental factors that impinge on the events of development or affect the adult complement of neurons and glia (Fig 1). For example, in development, the surrounding neuronal elements can invade the territory of other neurons by their expansion through growth. This can also be viewed as a passive genetic action. Other factors that impact neuronal form are physical and chemical interactions which affect the macromolecular organization.

X-irradiation and some kinds of chemicals influence genetic expression by reducing neurogenesis of specific neuronal types. These disproportions in neuronal types are also produced in conjunction with malnutrition, and viruses (See Hillman and Chen 1986 for references).

In determining factors responsible for the emergence of shape of neurons and their arbors, generation of specific pertubations with analysis of parameters of shape are useful tools to characterize factors as being either intrinsic or extrinsic determinants. Parameters that have minimal change are the least likely to be influenced by the environment and is most likely to have an intrinsic origin (Hillman 1979). Thus a useful approach is to analyze parameters for their invariance (Hillman and Chen 1984) following a series of perturbations. When a number of parameters change but a related parameter does not, the latter has a higher likelihood of being intrinsic. One such example is the constancy of total contact area of postsynaptic membrane specializations while the values for number and size of individual sites were altered reciprocally following various degrees of deafferentation (Hillman and Chen 1984).

2.0 PARAMETERS OF NEURONAL SHAPE The overall shape of neurons and their arbors have so far defied quantitative characterization due to a lack of concise descriptors for specific shapes of arbors. Nevertheless, the basis of global arbor shape can be defined in geometrical terms by parameters for size and orientation of segments in a relative coordinate system. Segmental parameters yield readily definable global parameters for: volume, surface area, and total length of processes (Fig. 2). Another major parameter of global arbor shape is branch pattern. This is represented by the connectivity of segments forming arbors and is definable as branching pattern irrespective of the arbor size or orientation.

A method of ordering segments of arbors for their position on the arbor now makes it possible to give a complete description of any neuron. Analysis of neuronal form by morphometric measures of segment parameters is currently the only effective means of establishing the rules for describing global shape of dendritic arbors. There are 3 principal categories for segment parameters: the relative size of segments (diameter and length), the relative orientation of segments to each other and to the soma, and the relative connectivity pattern of segments (branching pattern, Fig.2).

The size parameters of segments are represented by the relative scaling of length and diameter for all segments throughout the arbor. Diameter is best considered as cross sectional area for segments along the arbor. The length of segments can be expressed as the separation of bifurcation points or as the path length that processes follows.

The direction of segments originating from the soma and from bifurcations gives rise to global orientation of arbors around the soma. This domain is further defined by segment length and the direction of segments from bifurcation to bifurcation. The direction between bifurcations can vary as tortuosity that can be expressed as a difference in vectors between bifurcations as compared to the direction along the path that the segment encounters. The fundamental orientation parameter of segments, however, is best expressed in terms of the direction between the soma and the first bifurcation or between bifurcations.

The relative connectivity pattern of segments is the arrangement of segments in the arbor yielding global branching pattern of the arbor and has been called topology (Berry et al. 1965). Branching pattern can be defined by a method of terminal ordering of segments establishing connectivity relationships (Hillman et al. in preparation).

The relative size dimensions of the soma and the dendrites form a major part of neuronal shape. These together with descriptions of orientation and pattern of segments makes up essentially complete arbor representations that geometrically define neuronal shape. A working model of arbors can serve as a basis for determining parameters most likely involved in arbor shape.

3.0 NEURONAL ARBOR MODEL

An understanding of factors defining dendritic arbors globally can be gained from 5 concepts of arbors. The first is that the soma intrinsically defines a potential for emergence of a process or a number of processes. This potential is expressed as the specific amount of total cross sectional area for processes that can be generated from the soma (Fig. 3). The cross sectional area can be contained all in one process or divided between many of them (Fig 3).

A second part of the model is that each process emerging from the soma has a stem cross sectional area that is distributed by bifurcations of each segment and by their length until a limiting terminal segment diameter is reached (Fig. 3). This distribution of the cross sectional area as length generates the cylinder component of this model.

A third part of the model is that the cross sectional area can be altered in a number of ways along the course of processes from the soma to the terminals. First, the cross sectional area can increase or decrease along segments and is called segment taper (Fig. 4). Secondly, a taper may occur across bifurcations (Fig.4) and is classically known as branch power (Rall 1959). These two parameters markedly alter the number of bifurcations needed to reach a limiting terminal diameter.

A third parameter for cross sectional area is the distribution of the cross sectional area between the respective daughter segments called daughter-branch diameter ratio (Fig 5). The cross sectional area can be equal between the daughter segment or one can be as small as the cross

sectional area of a terminal segment. This parameter markedly alters the shape of neurons as a primary determinant of branching pattern of arbor shape. A final part of the model describes the orientation of processes. The projected direction of each stem segment arising from the soma establishes the direction of the arbor. At each bifurcation, the directions are altered and together with length generate a spread resulting in the spatial domain of the arbor.

Much of the difficulty in correlating models of neurons to actual neurons results from changes in cross sectional area between the stem and the terminals. Both length and diameter vary from segment to segment yielding patterns that are difficult to define. A means of ordering segments for establishing the variability along the segments composing the arbor is available as terminal ordering of segments (Hillman, Gelbfish and Chujo).

A method of placing order values on segments is useful to analyze arbors and correlate arbor shape with subcellular structure. Each successive level of segments from the terminals to the soma are labeled with values that represent the number of terminals that are distal along the arbor at any segment level. This method, called terminal ordering, is very useful for comparing segments of essentially the same cross sectional area for various segmental parameters. These order values also are used to define a numerical nomenclature of branching patterns providing a means to reconstruct branch patterns at a later time. Most importantly, the values can be applied to a scheme for classifying branch patterns according to their asymmetry and thus they establish up to 8 types of branching patterns. Neuronal classification can be generated from branch patterns and other types of parameters that numerically distinguish one type of neuron from another.

4.0 INTERRELATIONSHIPS BETWEEN SEGMENTAL PARAMETERS AND ARBOR SHAPE

An understanding of the contribution of segmental parameters to global arbor shape is essential for a meaningful conceptualization of neuronal form. Global parameters of arbors emerge from a combination of intrinsically determined segmental parameters and interactions with other neurons and glia. The shape of neurons can be further modified by epigenetic influences affecting genetically-defined interactions and sometimes even by intrinsic control. The latter occurs in conditions where specific neuronal components may be destroyed or fail to develop (Hillman and Chen 1984).

The emergence of global arbor parameters from segmental parameters is from a combination of intrinsic control and interactions which both contribute restrictions on dimensions that arbors can obtain. This produces a complexity that obscures the source of factors determining neuronal form. The determinants can subtly define aspects of shapes through their interactions that must be critically tested to distinguish intrinsic sources from extrinsic or epigenetic sources.

Analysis of the relative size parameters of component segments reveals strong interrelationships between segmental parameters in generating global shape of arbors. The parameters of diameter are: 1) stem segment diameter, 2) terminal segment diameters, 3) segment taper, 4) branch power, and 4) daughter branch ratio. Together these parameters express the distribution of cross sectional area from the soma to the terminals. Segment length is the principal length parameter which

interacts with diameter to yield the foundations for volume, surface area, total length, and branch pattern. The orientation of the stem segment from the soma and successive branch angles defines the orientation of the arbor. The interrelationship between orientation and size parameters are not readily definable.

4.1 Size Parameters of Dendritic Arbors

4.1.1 Stem Segment and Terminal Segment Diameter.

The generation of single or multiple stem segments from somas reflects an invariance that is related to soma size ( ). This is seen in the relationship between the combined trunk cross sectional area and the diameter of the soma. A strict relationship is seen in comparing a number of different neuronal types (Fig. 7). A plot of soma diameter to that for the combined-trunk cross sectional area generates a definative 1:1 slope for motoneurons alone. Pyramidal cells and granule cells fall on the same line as motoneuron. Purkinje cells have the same slope but have an offset representing a difference in gain.

The meaning of this soma diameter to the combined-trunk cross sectional area relationship remains obscure but may represent a coupling of single dimensional element of the soma to the cross sectional area of the processes. The logical dimensional relationship is the number of some substructures that generate the combined-total cross sectional area for all processes arising from the soma. The likely structures are Neurotubules (see below). At another dimensional level, the volume of the arbor is related to the area of soma structure such as the surface of

the nucleus, plasma membrane, endoplasmic reticulum, etc. Thus, the total amount of cross sectional area that a neuron can produce is limited by the active area of some subcellular structure.

The cross sectional area of stem segments represents much of the extent of the arbor since this area must be distributed to terminals. Therefore, the number of bifurcations on the arbor reflects the cross sectional area of the stem segment. Furthermore, the cross sectional area of terminal segments influences the total number of bifurcations in the arbor by specifying whether or not a bifurcation can occur. Thus, stem and terminal diameters determine the size of the arbor by limiting the number of segments that make up the arbor.

4.1.2 Branch Power and Segment Taper

Two cross sectional area parameters, branch power and segment taper, alter the relative size of the arbor by affecting the distribution of cross sectional area across bifurcations and along segments lengths (Hillman 1979). Both represent parameters of taper and can be positive or negative. Positive taper results in larger branch patterns because more bifurcations are needed to reach the limiting terminal diameter. For example, the numerous terminals generated in Purkinje cells result from a positive taper for segments and for bifurcations. Arbors with large stem segments such as motoneurons have few bifurcations because of a large negative taper. This occurs along segments and to a degree at bifurcations. The amount of sharply localized change in cross sectional area occurs the bifurcation as compared to the segment taper over the measureable length involving the bifurcation.

The most marked segment taper is seen in the stem segment and in terminal segments. The latter may occur because limits for bifuracation are met but the process continues to extend for some distance and looses some of its substructure.

4.1.3 Branch Pattern and Diameter of Daughter Branches.

The analysis of segmental parameters shows that branch pattern of arbors is highly related to the ratio of cross sectional area distributed between the pairs of daughter branches. The pattern of branching can be near symmetrical but often has a degree of asymmetry that ranges to nearly maximum. The maximum asymmetry is produced when the ratio of the daughter branches are the greatest, ie. the size of the smaller segment is equal to a terminal branch and the side branches can no longer bifurcate. Equal distribution of cross sectional area allows continued bifurcation to the same amount for all branches. The result is a symmetry of the arbor. As the ratio becomes more asymmetrical, the branch pattern is more asymmetrical. Intermediate patterns of branching occur from intermediate patterns of cross sectional area symmetry and from a mixture of ratios from the soma to the terminals. These patterns have distinct functional implications as based on cable properties of dendrites. Branch pattern has been impossible to describe because of a lack of methods specifying the order of segments along the arbor.

A classification of branching pattern was recently specified by Pelter and Wehr based on ambilaterality ( ). An alternative method using terminal segment ordering (Hillman 1986) provides a means of ranking arbors, naming them, and classifying types of arbor patterns as based on symmetry. Using the terminal ordering method, a symmetry index can be generated for any arbor. When the symmetry index values are applied to a numerical scheme for classifying arbor types, the amount of branch pattern asymmetry is defined and the arbors can be placed in one of four or eight classes of branch patterns.

4.1.4 Segment Length.

Segment length varies considerably though each neuronal type has a range for each segment level between the soma and terminals. For example, Purkinje cell terminal segments are uniformly short as are the claws forming granule cells. On the other hand, pyramidal cells have very long terminal segments and may continue to grow in aging (Buell and Coleman). Length of segments separates the bifurcation points and interacts with cross sectional area as a component of taper. Length is also a major counterpart of orientation for indivdual segments. The extent of the arbor as total length is generated exclusively through length while the volume and surface area of arbors are formed in combination with cross sectional area. The distribution of length for segments of various cross sectional areas can be a major factor in regulating surface area and volume relationships.

The intrinsic control of length is unlikely produced at the segmental level. Rather, length is indirectly controlled as total length through volume of the arbor occurring as an interrelationship between cross sectional area and length. Most evidence points to total arbor volume of each neuron as being controlled intrinsically. This is logically a product of subcellular structure (see below).

4.2 Orientation.

The orientation at the segmental level is represented by three major factors: direction of the two ends of segments emerging from the soma, direction between segment ends from the parent segment, and change in direction between the two ends of the segments as tortuosity. A parameter for the latter can be obtained by comparing path length with vector length between the ends of the segments.

Orientation of segments is determined largely by interactions with other cellular elements. There are, nevertheless, predisposing forces that are dictated by each neuron. These are seen as the tendency for side processes of growth to emerge at right angles to other processes and emergence of the parallel fiber axon from the two opposite ends of the premigratory neuron.

5.0 RELATIONSHIPS BETWEEN PARAMETERS OF ARBOR SHAPE AND SUBCELLULAR STRUCTURE

Substructural elements (in conjuction with extrinsic factors) play major roles in defining branch patterns and the relative size of segments. Microtubules are best known for their role in dendritic shape while neurofilaments are the major constituents of axons. Additionally, internal membrane structures, endoplasmic reticulum and mitochondria influence neuronal shape as less regular elements that add volume and indirectly augment shape. Together with associated proteins and cytoplasmic fluid, these subcellular components form segments giving rise to arbors through bifurcation to form patterns. These subcellular components are the skeleton for a number of segmental parameters whch together generate defined global parameters of arbor shape.

5.1 Neurotubules.

Neurotubules form a base for process extension as a skeleton that provides stability to length (Yamada and Spooner). The longitudinal orientation of neurotubules and their cross sectional spacing suggest that the neurotubules form the axial core of process development and long term stability. Thus, neurotubules form the volume core of dendritic arbors by generating the cylinder-basis of processes. The distribution of neurotubules through bifurcation and length affects volume distribution along the arbor, thus affecting the surface area of neurons globally.

Reconstruction of microtubules indicates that they are continuous rather than consisting of short segments telescoping in dendrites (Table 1). Microtubules can be considered to begin in or near the soma and extend to or near terminals. A few may have beginnings away from the soma or end before a final bifurcation 9Fig). In dendrites, the microtubules have relatively uniform patterns with exception of spiny dendrites of Purkinje cells and the predominante neurons of the striatum. The loose pattern has no regular arrangement and the density is lower due to numerous mitochondria and in some cases prominent endoplasmic reticulum. In large dendritic segments without taper, neurotubules density is closely maintained throughout the cross sections of a large range of segment sizes. The relatively constant density of neurotubules in dendritic arbors of various sizes in the absence of filaments reveals that tubules are the principal elements defining the cross sectional area of all dendritic arbors (Filaments can add to this dimension; see below).

Analysis of the stem segments of Purkinje cells (having only one dendrite emerging from the soma) reveals that a consistent maximum number of neurotubules emerges from the soma. This is seen as a plateauing of the neurotubule number in dendritic cross sections as the Purkinje cell soma is reached (Fig.). Each neuronal soma appears to have a definative number of neurotubules which form the core for cross sectional area of processes as they emerge from the soma.

The distribution of the total number of neurotubules arising at the soma and extendind into a number of processes forms the basis of multipolar neurons. These can emerge in different directions from the soma to give an array of soma-arbor patterns.

At bifurcations, the microtubules are clearly separated into bundles having sizes that match the cross sectional area of each branch (Hillman 1979). Although a few neurotubules may continue for a short distance within the larger straighter segments at bifurcations, they soon loop back and exit into side branches. This strongly suggests that neurotubules play a role in determining the ratio of cross sectional area (daughter-branch diameter-ratio, DBDR; daughter-branch cross sectional-area ratio, DBCSAR) at bifurcations. Therefore, since the branching pattern of the arbors is determined by the ratio of the cross sectional area between the two arbors, neurotubules must have a role in branching pattern. The actual determination of the bifurcation site and the dividing up of the cross sectional area (directing of the neurotubules into the appropriate process) is an interactive event. Nevertheless, the force to bifurcate appears to be intrinsic with generation of filopodia as pilot elements that test the immediate surrounding region for interactive elements. The selection of these filopodia is then made interactively.

Another major role of neurotubules in determining arbor shape is a limitation on the diameter of terminal segments. The minimum number of neurotubules that can be divided between two segments appears to be in the 12 to 14 range. This limit produces terminal segments that contain a range of neurotubules from 3-4 up to 12 or13 (Fig.). Those terminal segments that have less than 10 either were subject to a minimum number at the final bifurcation or have lost neurotubles as is evident from the marked taper in long terminal segments. The diameter of terminal segments just after the final bifurcation point is, thus, highly determined by intrinsic factors limiting the minimum number of neurotubules.

An example of such limitation on neurotubule number is seen in the clustering of neurotubules when quantitating them in terminal segments of Purkinje and pyramidal cells. A maximum of 12-14 neurotubules appear to be the optimum number at the base of the final bifurcation. Fewer occur along the course of terminal segments, however, they would represent the number obtained from bifurcations of the 24-28 neurotubule group. A taper over the length of terminal segments is common and adds to this decrease yielding some areas of terminal dendritic segments with as few as three neurotubules. For example, dendrites can lengthen with aging (Buell and Colman 198). Currently it appears that 3 tubules constitute a viable limit for dendritic processes.

5.2 Neurofilaments.

In neurons that have much taper along their processes, neurofilaments interact with the regular pattern of microtubules to expand the cross sectional area at the base of the arbor. The gradual loss of the filaments or their colessing to fewer number alters the density of microtubules between the base of the arbor and the more distal dendritic segments (Hillman 1979). This rapidly reduces the expanded cross sectional area from the base, distally, so that fewer bifurcations are needed to reach the limiting terminal diameter. The change in total cross sectional area between the soma and the terminals occurs as segment taper and as branch power less than 2/2. It must be assummed that the associated proteins of neurotubules and neurofilaments are additional elements that support cross sectional area and may be also a major factor in taper along segments.

5.3 Endoplasmic Reticulum and Mitochondria.

The amount of cross sectional area capable of being generated by neurotubules and neurofilaments and their associated proteins is augmented by the presence of endoplasmic reticulum and mitochondria. In Purkinje cells, a combination of subsurface membrane reticulum and lattice sheets of reticulum perforated by neurotubules form a major constituent of the main dendritic arbor. The reticulated network channel follows the plasma membrane to form a cylinder that is bridged across the diameter by a continuous lattice of membrane channels. The density of this reticulum is so extensive that some sheets of the membrane channel appear perforated by microtubules since the lie within the minimum distance of the channel enclosing membranes.

In spiny branchlets, mitochondria add considerably to the cytoplasmic mass, however, the lattice organization is lost. The endoplasmic reticulum extends from the subsurface reticulum to into the spines to form a complex folding within the head of the spine. In pyramidal cells this has been called the spinous apparatus.

6.0 INTRINSIC VERSUS EXTRINSIC CONTROL OF NEURONAL FORM.

The intrinsic control of neuronal arbor shape is evident in only three segmental parameters and in one global parameter of arbors. These segmental parameters are limited to cross sectional area and are: total amount of cross sectional area of the stem segment, limiting cross sectional area of individual terminal segments, and taper along segments. The generation of the 3 parameters is primarily from substructures, neurotubules and neurofilaments which constrain the cross sectional dimension of arbor segments.

The principal subcellular effect appears to arise from the total number of neurotubules and intermediate neurofilaments together with asssociated proteins that space these structures. These alone appear to generates a relatively predictable cross sectional area for each type of neuron. This invariance in cross sectional area related to the number of neurotubiules arising from the soma is carried into each of the stems of arbors. However, the dividing up of this maximum number of neurotubules into individual arbors is not determine intrinsically by the neuron. Stem segment diameter is largely a product of interactive forces that control process growth through directing of neurotubules into selected filopodia arisiing from the soma during development. The final number of neurotubules at the terminal bifurcations appears to be a limit to the ability of interactive forces to divide up a small number of neurotubules. The final number of neurotubules to support a process appear to be intrinsically determined by a limit of three neurotubules.

Taper is generated when large neurofilaments are present. These intrinsically determined structures may in combination with some asssociated proteins generate an enlargement of the arbor segments near the base.

The total amount of assembled tubulin may may control volume through cross sectional area in combination with segment length. The total length of assembled tubulin as segments arranged in dendritic processes may be intrinsically determined. This would explain strict relationships between the size of the soma and the volume of the arbor.

Much of arbor shape is determined by extrinsic interactions with processes of other neurons as well as by epigenetic influences from the surrounding neuropile (fig. 1. Cross sectional area of individual stem segments arising from the soma and ratio of cross sectional size between daughter segments are interactively controlled.

The intrinsic control of length is unlikely at the segmental level, however, most conceivably length is indirectly controlled as total length through volume of the arbor. This occurs in the interaction between cross sectional area and length to yield volume. Most evidence points to total arbor volume of each neuron as being controlled intrinsically. This is most logical from the stand point of subcellular structure.

Surface area of neurons is generated by the distribution of the volume through positioning of segment length along various segments having larger or smaller cross sectional areas. Length of segments is primarily determined by interactive factors. Neuronal masses and fiber bundles form barriers that influence shape of the arbors in a passive mode. Other barriers are the vasculature and brain surfaces formed by pia and ependyma. The physical presence of somas and processes of both neurons and glia contribute to tortuosity of processes and some case postions of entire groups of neurons.

SUMMARY

An understanding of intrinsic and extrinsic (interactive and epigenetic) control of neuronal form requires correlations between substructure and shape parameters for neuronal somas, dendrites and axons. Descriptors of neuronal shape are the size of the soma and the size, orientation and patterning of segments making up the axonic and dendritic arbors. The minimum essential parameters for describing the soma-dendritic pole of neurons are the diameter of the soma, the total cross sectional area of all dendritic segments emerging from the soma, the cross sectional areas of individual stem segments of arbors, the limiting cross sectional area restricting further bifurcation, and segment length.

Additionally, the shape and size can be further elaborated by changes in cross sectional area occurring between the soma and the terminations of final segments for each bifurcation path. These size changes are taper (segment taper and branch power), and the ratio of cross sectional area distributed to the two daughter branches. The length of segments can be expressed in two useful parameters, bifurcation to bifurcation distance and segment path length. The orientation of segments from the soma to the first bifurcation and between bifurcations provides the principle directions of the arbor with a tortuosity parameter expressing deviations that the segment takes between bifurcations. The complexity of the arbor is largely the result of variations in parameters for length, taper, and ratio of cross sectional area for the various segments. A complete description requires a means of labeling the segments so that the variations can be properly defined for all parts of the arbor.

Based on the variability of these parameters in various neuronal types one can conclude that neuronal shape parameters are intrinsically controlled by the neuron through: 1) total cross sectional area of stem dendritic segments, 2) a minimum cross sectional area of terminal dendritic segments, 3) taper of dendritic arbors occurring along segments and across bifurcations, and 4) total volume of dendritic arbors of neurons.

The subcellular structures that are most responsible for these parameters are neurotubules and neurofilaments. In the stem segments of arbors the invariance in the total number of neurotubules projecting from the soma define the core of a relatively invariant total cross sectional area of the combined stem segments. The ability of neurotubules to distribute into two smaller bundles, until they reach a limitation on their number, repesents the basis of the bifurcation principle. The result is that bases of terminal segments are in a range of diameter that supports a neurotubule core from 12-14 down to 3- 4 neurotubules. Further reductions from 12 to three tubules appear to occur with loss of tubules and results in taper commonly found in terminal segments.

Some neuronal types have bundles of intermediate neurofilaments in their dendrites. These cells also have a taper along the segments and across bifurcations. Intermediate neurofilaments appear to be altering the distribution of cross sectional area from the base of the dendritic arbor to terminals by displacing the relatively regular arrangement of neurotubules in the segments near the base of the arbor. In distal segments, the filaments disperse among the neurotubules and the density of neurotubules increases.

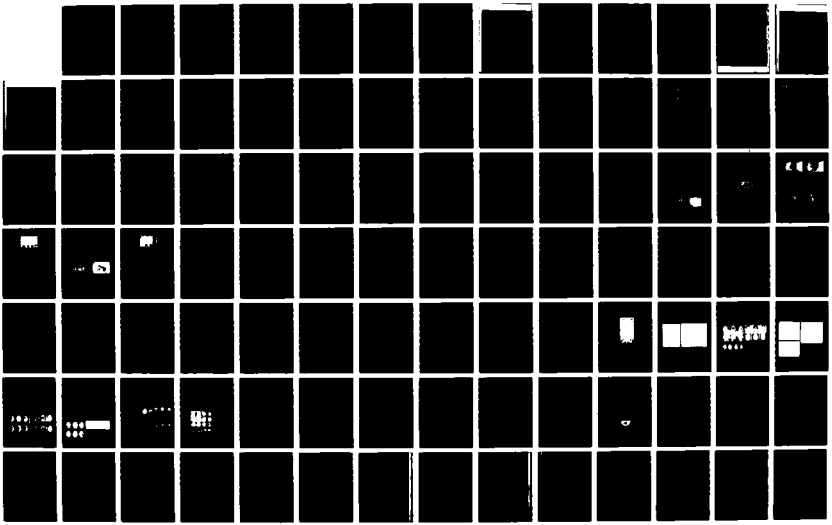
A fourth intrinsic affect is related to the volume of the arbors. The total volume of dendritic arbors is closely correlated with the size of the soma. In dendrites, the core for this volume appears to be determined by the amount of assembled tubulin in conjunction with spacing produced by associated proteins that generates the neurotubule spacing.

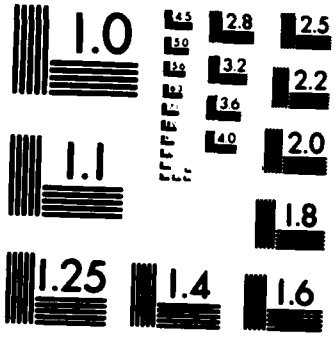
The intrinsic determination of neuronal shape is also influenced by other subcellular organelles such as ER and mitochondria. The axial stabilizing effect of the neurotubules and associated proteins serves as a pliable skeleton on which filaments and aother elements may be interspersed if present.

The ER interlaces along the axis of this skeleton as continuous channels adding to the cross sectional component. A major channel follows the plasma membranes and is bridged across the diameter by channels that interconnect as both a axial and coranal lace. Elaborations of the cross sectional area of processes occurs as spines which are composed of an extension of the continuous endoplasmic reticulum channel and a microfilament matrix. Mitochondria are interspersed along the axis in varying densities but do not enter true spines.

AD-A188 809 PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP HELD IN COLLEGE..(U) TEXAS A AND M UNIV COLLEGE STATION R B LIVINGSTON AUG 85 4/5

UNCLASSIFIED DAAD-17-85-G-5042 F/G 6/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

FIGURES

Fig. 1. Diagram showing a proposed relationship between intrinsic, extrinsic and epigentic factors affecting neuronal form. The intrinsic influence is defined largely by the genetic codeon expressed through subcellular structure. The extrinsic control is genetically expressed through interactions with other cellular components. Passive or indirect genetic expressions also occur as barriers that alter the spatial domain. Epigenetic influences from the environment further affect the interactive expression through sensory input and direct effects on the cellular elements such as by preventing development or eliminating cellular components. Interactions may be affected by direct effect on chemical constituents necessary for interactions.

Fig. 2. Global parameters of dendritic shape and their derivation from segments. The 3 major parameters of segments are cross sectional area, length and orientation. These can be defined in a general coordinate system which completely describes the arbor in a global sense. A set nof global parmetes that completely describe arbors has not yet been defined.

Fig. 3 Diagram showing generation of neuronal processes from a soma as total cross sectional area and its distribution to terminals by bifurcation. In this model, cross sectional area for generating dendritic processes emerges from the soma as a single process or is divided between a number of processes. In either case the total cross sectional area is constant. The distribution of the cross sectional area to the terminals is made by bifurcation. The terminals have a limiting size that resticts further bifurcations. The total amount of cross sectional area reaching terminals is the same if taper modifications does not occur along their course.

Fig. 4. Arbor taper represents changes that can occur in the cross sectional area between the stem segment and the terminal segments. This may involve length as a segment taper or possibly sharp increments at bifurcations and is defined by branch power. The cross sectional area can become larger (positive taper) or it can decrease (negative taper). Both positive and negative tapers can occur on the same dendrite.

Fig. 5. A second major variation in cross sectional area occurs at bifurcations as a differential in distribution of this area between the two daughter segments. The degree of this differential alters the number of bifurcations that are needed along each path from the soma to terminals for attaining the limiting diameter.

FACTORS INFLUENCING NEURONAL SHAPE & ORGANIZATION OF CIRCUITRY

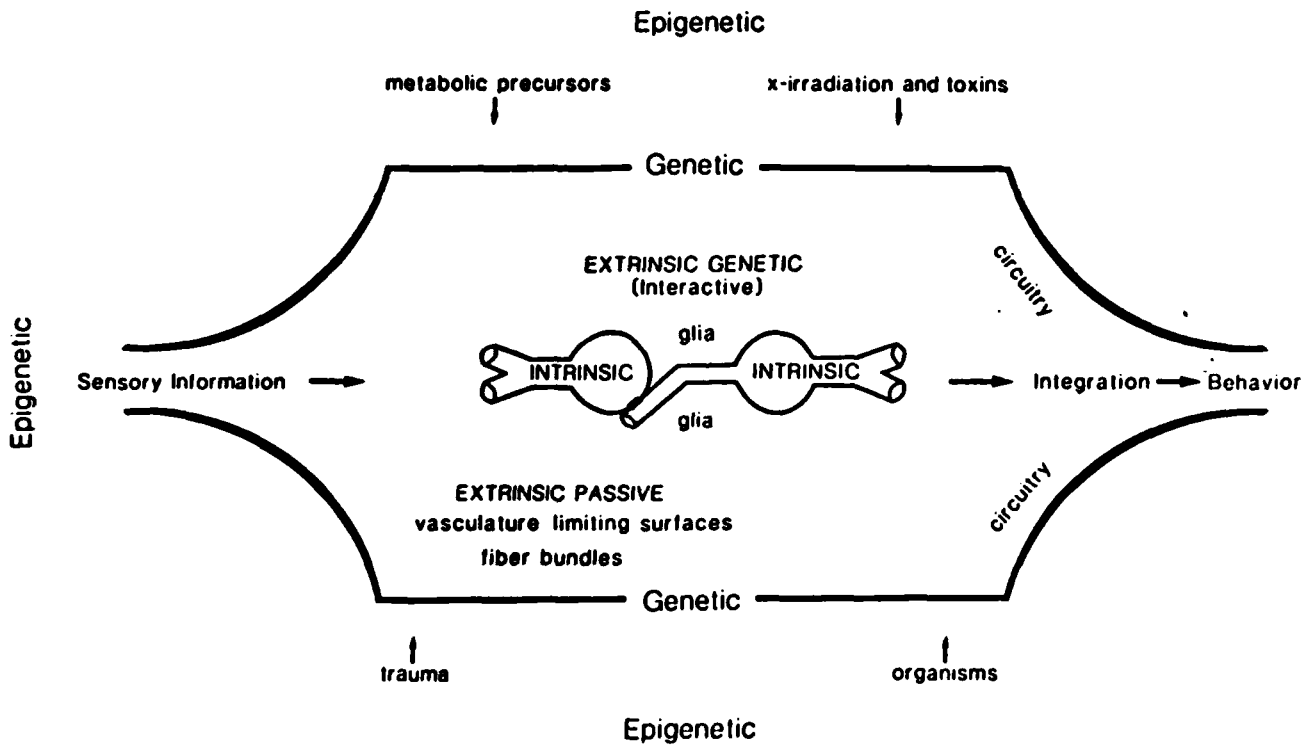


Figure 1

PARAMETERS OF DENDRITIC SHAPE

SEGMENTAL	GLOBAL	
A. Diameter (Cross sectional area) a. Stem segment diameter b. Terminal segment diameter c. Branch power d. Segment taper e. Daughter branch diameter ratio	A. Arbor Cross Sectional Area (STEM, TERMINALS & TAPER)	Branch Pattern Total Volume Total Surface Area
B. Length a. Segment length (Path) b. Distance between bifurcations	B. Total Length	Spatial Domain
C. Orientation a. Segment orientation From soma From bifurcation Tortuosity	C. Orientation of Arbors	

Figure 2

3.8.30

DISTRIBUTION OF TOTAL CROSS SECTIONAL AREA INTO DENDRITIC SEGMENTS

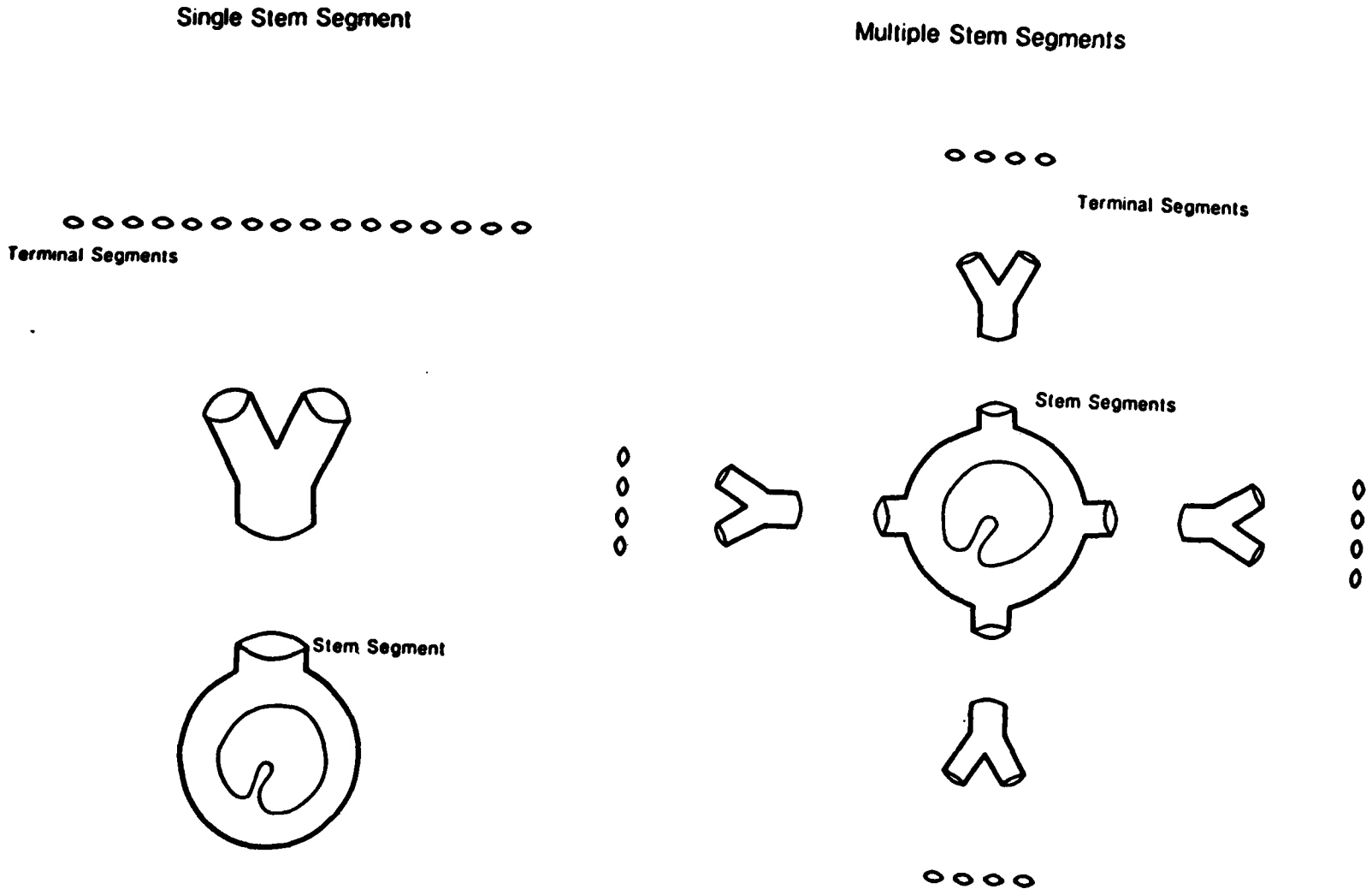


Figure 3

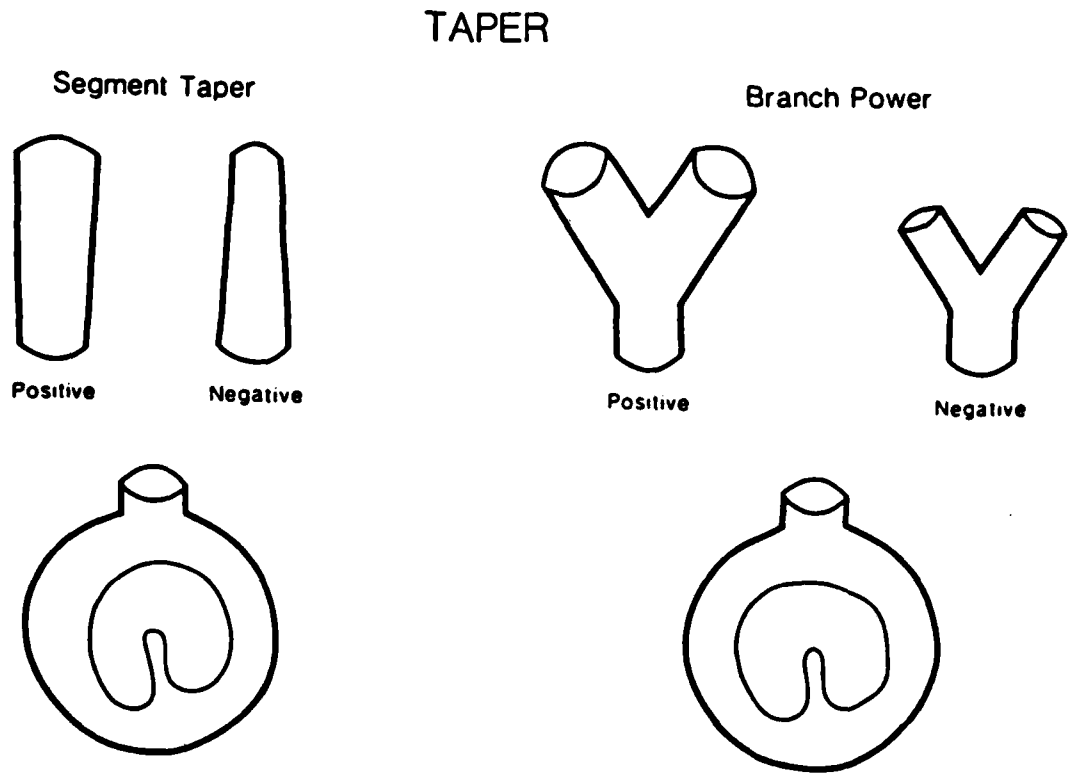


Figure 4

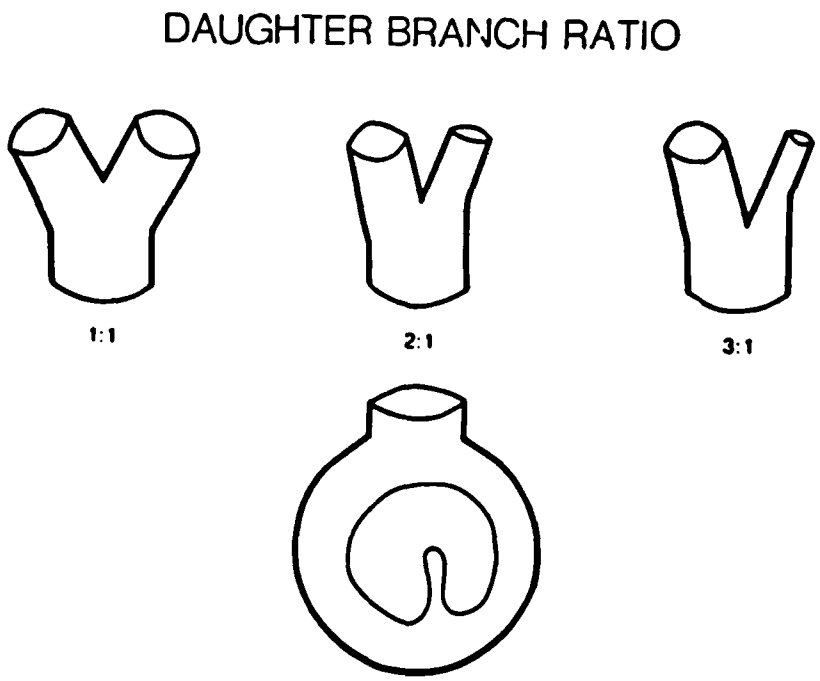


Figure 5

SEGMENT LENGTH

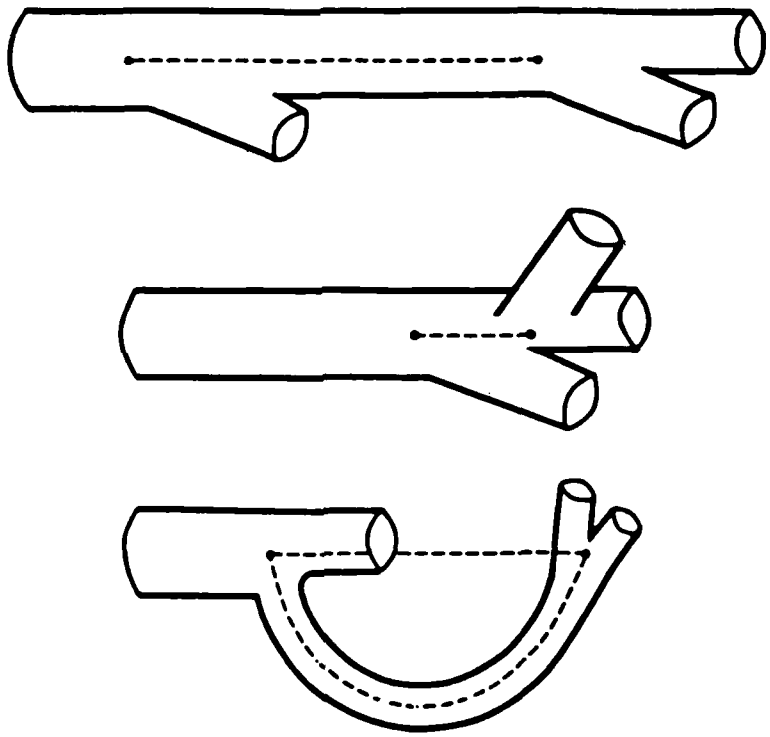


Figure 6

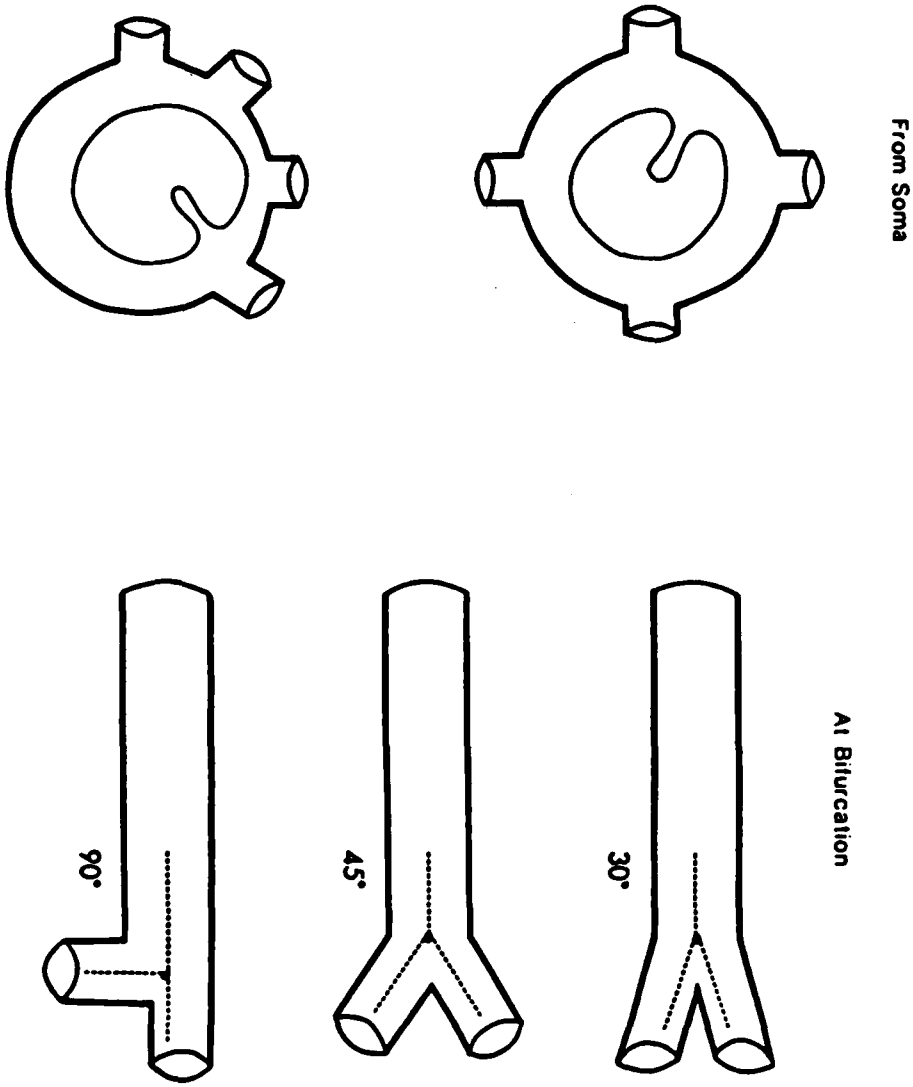


Figure 7

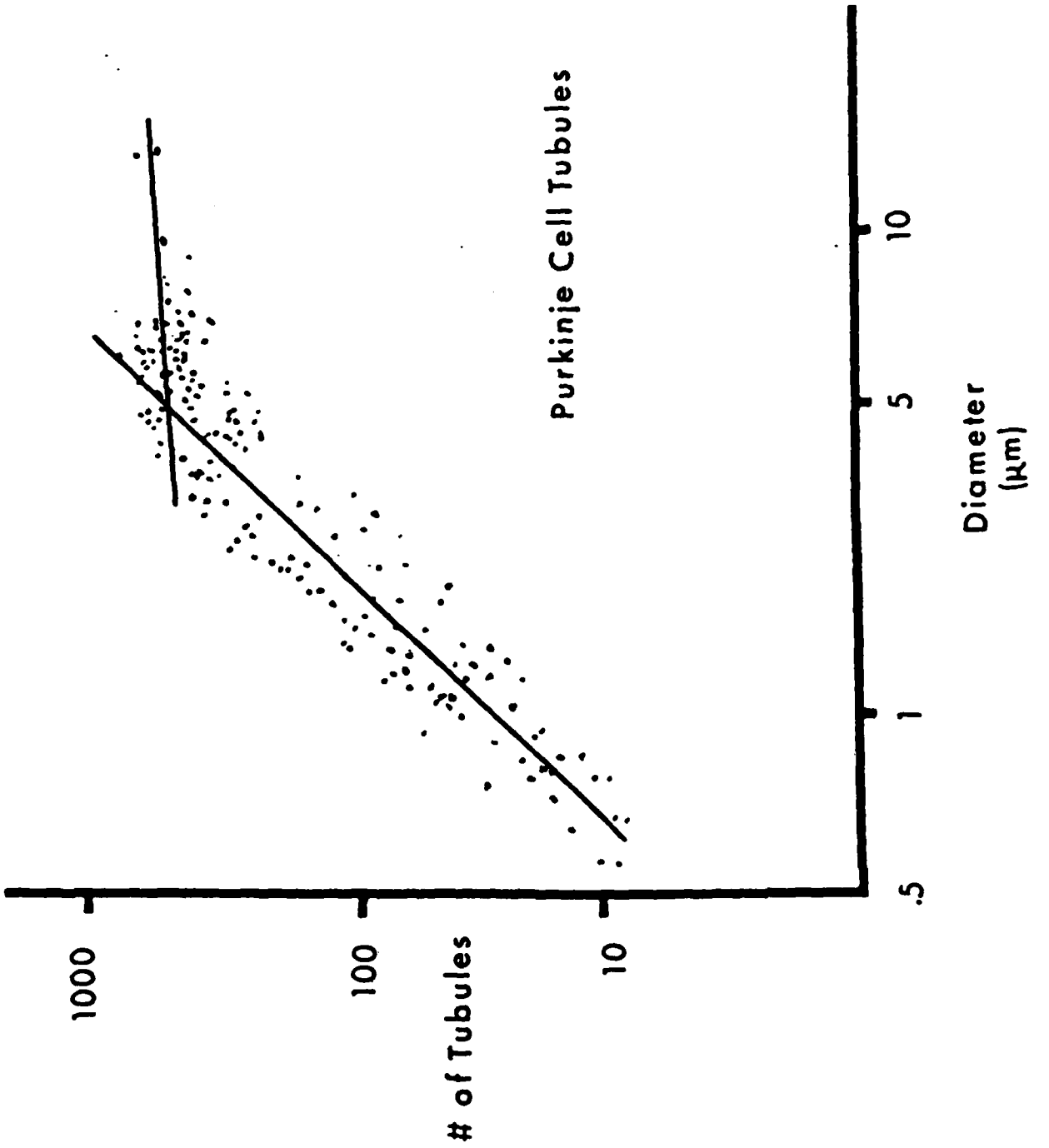

Purkinje Cell Tubules
# of Tubules
1000
100
10
.5
1
5
10
Diameter
(μm)

Figure 8

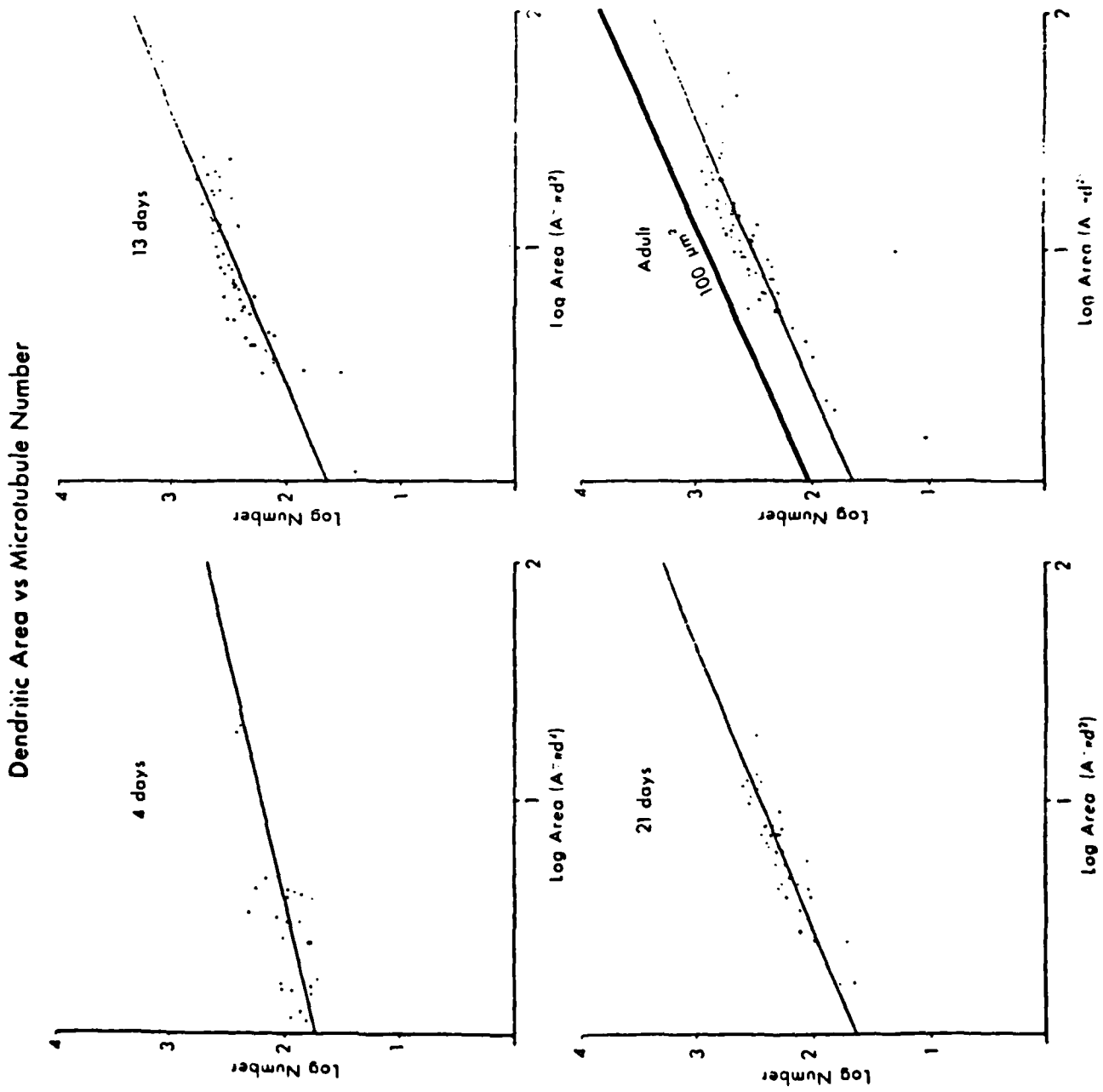

Dendritic Area vs Microtubule Number
4 days
13 days
21 days
Adult
100 μm²
Log Number
Log Area (A~πd²)
1
2
3
4

Figure 9

CONTROL OVER SEGMENTAL PARAMETERS OF DENDRITIC SHAPE

Intrinsic

DIAMETER :

- Stem Diameter
- Branch Power
- Segment Taper
- Terminal Diameter

Extrinsic

DIAMETER :

- Stem Segment Diameter Distribution Over Soma
- Daughter Branch Diameter Ratio

LENGTH :

- Segment Length

ORIENTATION :

- Stem Direction From Soma
- Segment Direction From Mother Segment

Figure 10

NEURONAL SHAPE.
DENDRITIC PARAMETERS & SUBCELLULAR DETERMINANTS

PARAMETERS		DETERMINANTS
Soma Shape		**Subcellular Structures**
Size		Nucleus . ER . Filaments . Tubules .
Total Arbor Amount		Total # Tubules
Arbor Number		Distributing Tubules to Arbors
Dendrite Shape	**Segment Parameters**	**Subcellular Structures**
Size	Stem Diameter	Total # Tubules & Filaments
	Terminal Diameter	Minimum # Tubules
	Taper	Filament Distribution
	Daughter Diameter Ratio	Tubule Distribution
	Segment Length	
Branching Pattern	Daughter Diameter Ratio	Tubule & Filament Distribution to Segments
Orientation	Soma to Stem	
	Parent to Branch	
	Tortuosity	

Figure 11

INTRINSIC DETERMINANTS OF DENDRITE SHAPE

Number of Neurotubules Generated by Neuron	→	Total Dendritic Cross Sectional Area
Minimum Number of Neurotubules to Bifurcate	→	Minimum Diameter at Terminal Segment Base
Neurofilament Number & Distribution	→	Segment Taper & Branch Power
Total Number of Assembled Tubulin Molecules	→	Total Volume of Dendritic Arbors

INTRINSIC FORCES INFLUENCING ARBOR SHAPE

Formation of Numerous Filopodia	→	Generates 20-30um search area
Substructural Invasion of Extrinsically Selected Filopodia	→	Generates Dendritic Segment
	→	Generates Bifurcations

Figure 12

In: The Neurosciences: Fourth Study Program. Edited by F.O. Schmitt and F.G. Worden. MIT Press: Cambridge, Massachusetts and London, 1979.

27 Neuronal Shape Parameters and Substructures as a Basis of Neuronal Form

D. E. HILLMAN

ABSTRACT Computer technology has provided a means to record neuronal arborizations in three dimensions and to analyze the data for fundamental parameters of neuronal shape. Quantitative values for the elements comprising the cytoskeleton correlate with the parameters of shape and subcellular structure.

Seven parameters are needed to define the fundamental aspects of dendritic shape. The basis for this assumption is that the cross-sectional area of a process that emerges from the soma is divided successively by branching until a limiting terminal diameter is reached. Branch points serve to define segment length, branch power, daughter-branch ratio, and segment orientation. Tapering may occur between branch points.

Underlying neuronal shape are prominent structural components of arborizations (tubules and filaments) comprising a cytoskeleton that may be involved in maintaining the diameter of dendrites. This is evident in some neuronal types where microtubule density is constant throughout the arbor and the branch power equals two (cross-sectional area is preserved along the extent of the arbor). In motoneurons the branch power equals 3/2. Here both filaments and tubules shape the dendritic processes. Filaments produce a geometrical tapering of the tree (beginning at the soma, the cross-sectional area is decreased at and between branch points).

The determinants of form are discussed in relation to the variability of the fundamental parameters. It is shown that stem diameter, segment length, daughter-branch ratio, and orientation are considerably variable and thus are most likely determined through interactive influences during development. Total stem diameters arising from soma, branch power, segment taper, and terminal diameters are much more tightly constrained and may be largely determined by the intrinsic influences acting through the subcellular components of the cell.

Introduction

OUR PRESENT concepts regarding neuronal form have evolved over the last 150 years and have been largely dependent on the development of microscopic techniques. In particular, new staining methods and refinements in light optics ushered in a morphological revolution in the latter part of the nineteenth century. At this time, using cerebellar tissue, neuronal cell bodies were first seen with the light microscope (Purkinje, 1837). Later, groups of cell bodies were found in various regions (Deiters, 1865). Although the subsequent visualization of a network of processes lacing throughout the brain (Schulze, 1871) was instrumental to the development of a concept of the form of neurons, the key development was the discovery of the Golgi technique, which first allowed visualization of individual neuronal somata together with their processes (Golgi, 1874).

An early concept, held by the reticularists, viewed the nervous system as composed of a syncytium of cells and fibers. This concept was largely due to the development of the reduced silver reaction, which revealed within the cell processes fibrillar material that appeared to form a continuous meshwork connecting the various neuronal elements (Apáthy, 1897; Bielschowsky, 1902).

It was not until the Golgi technique had been applied to numerous areas of the nervous system that a compelling case was made for the existence of neurons as independent cellular elements (Ramón y Cajal, 1909, 1911). This view was initially stated by Waldeyer (1891) and Kölliker (1891) as the neuron doctrine. The strongest support for this concept came from the work of Ramón y Cajal (1911), who demonstrated not only that the nervous system was composed of several cell types, but that particular types of neurons were characteristic of given regions of the nervous system. This work also revealed that the meshwork previously described actually consisted of an overlapping of many orderly, treelike processes emerging from the cell bodies.

D. E. HILLMAN Department of Physiology and Biophysics, New York University Medical Center, New York, NY 10016

The final demonstration that the nervous system was composed of individual neurons and was not a continuous network came with the advent of the electron microscope. The high resolution of this instrument revealed that each neuron was a separate cellular entity and that "contact sites" were actually separated by extracellular space or synaptic clefts without continuity of the neurofibrillar material (Palay, 1956). Throughout this history, a vast amount of information has been collected regarding the different neuronal processes and the circuits they generate.

Neurons and neuronal circuits were initially understood and displayed in drawings by such people as Purkinje (1837), Golgi (1874), Kölliker (1891), Ramón y Cajal (1909, 1911), and Lorente de Nó (1934). Although drawings remain a principal way to depict neurons, this method has obvious drawbacks (Figure 1). For example, a true three-dimensional image cannot be obtained, and the size of the cells can only be approximated. Methods have therefore been developed to describe nerve cells quantitatively (Bok, 1936a; Sholl, 1953) and to display these constructs in three dimensions (Mannen, 1966).

An important advance along these lines was made when computer technology was applied to the analysis of neuronal morphology (Glaser and Van der Loos, 1965). Many nerve cells could be reconstructed in the computer and rotated and displayed from any viewpoint; dissimilarities as well as consistencies in the shape of neurons could then be fully appreciated. Even here, however, the need to view neuronal circuits in three dimensions—as in holograms—became apparent. Thus, as the tools we have for "seeing" neurons improve, what we see changes because our concept of neuronal form changes.

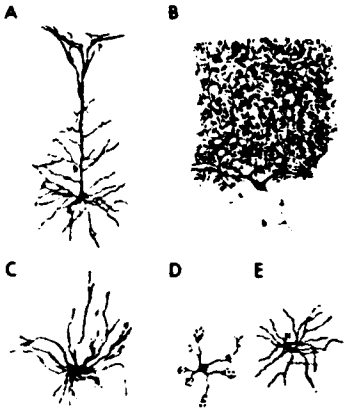


FIGURE 1 Illustrations of neuronal dendritic arborizations by Ramón y Cajal. (A) Pyramidal cells are multipolar with a single apical dendrite (b), which gives rise to branching patterns that differ from basal arbors (a). (B) The human Purkinje cell has two major trunks (other mammalian forms usually have only one). An elaborate arborization that spreads in a single plane through the molecular layer of the cerebellum has more numerous terminals and branch points than pyramidal cells. (C) Motoneuron dendrites have large stems and few branch points. (D) The small cerebellar granule cell usually has only 3–6 dendrites, which end in clawlike formations. (E) The stellate cell has long terminal dendritic segments. (From Ramon y Cajal, 1909: Figures 8, 9, and 129; Ramón y Cajal, 1911: Figures 23 and 24.)

Although the term "form" is often used by morphologists, what is usually meant is "shape." However, in this paper "shape" is used in a geometric sense and refers to the three-dimensional outline of each cell including its processes. Shape is independent of content and size. "Form" is used in an architectural sense and refers to the combination of shape with structure (form is also independent of size). I introduce these definitions here because, with the refinement of techniques, the quantification of neuronal shapes and their underlying cytoskeletal elements is now possible; thus, for the first time, we can truly speak of neuronal *form*. Developmental studies (Rakic, 1974; Berry and Bradley, 1976b) show, however, that the shape of nerve cells is influenced not only by intracellular structural elements but also by the environs of the neurons during development; that is, a study of neuronal form must consider ontogenetic influences.

Thus, with a quantitative description of neurons in hand, we can proceed to explore both the ultrastructural and the developmental basis for cell shape. Following this thought, this paper has been organized in three parts. First, from the mass of quantitative information available on cell shape, the fundamental parameters of form are extracted (i.e., those measurements essential to providing a complete description of a neuron). Next, the number and distribution of the subcellular structures underlying these parameters are determined. Finally, this information is used to establish which factors in ontogeny determine the values of the various parameters, utilizing a quantitative determination of variance within the parameters. The amount of *variability* is used as a key, together with the vast amount of information concerning the developmental process, to establish (although tentatively) which of these parameters may be determined by intrinsic and which by interactive (extrinsic) factors.

Quantitation of neuronal morphology

Although the polymorphic character of neuronal arborizations has long been known, the quantitative

analysis of these elements has been attempted only recently. Among the first to develop techniques for the quantitation of neurons were Bok (1936a) and Sholl (1953). These early methods were modified and used extensively in studies concerning the effects of various perturbations on neuronal shape (Eayrs, 1955; Clendinnen and Eayrs, 1961; Schadé and Groenigen, 1961; Coleman and Riesen, 1968; Schadé and Caveness, 1968) and have since been altered to eliminate certain artifactual results (Colon and Smit, 1970, 1971; Berry, Anderson, Hollingworth, and Flinn, 1972). A recent advance has been the use of computer methods to record and analyze neurons; following the pioneering work of Glaser and Van der Loos (1965), these methods have been elaborated and a number of approaches developed (see Brown, 1976; Lindsay 1977a). Network analysis, an entirely different and very fruitful approach, has been used by Berry et al. (1975) to define branch patterns and their progressive change (e.g., during development).

The application of these techniques has generated a large literature in which neuronal shape has been measured in multiple ways. In an attempt to unify some of these results, a set of parameters has been established. The criteria are that each parameter must (1) be easily measurable, (2) be consistent with and meaningful to the development of electrophysiological models, (3) be useful for comparing different types of arborizations, and (4) be nonredundant. Finally, the full set of parameters must be sufficient to describe the shape of neurons. These parameters would then serve as a fundamental basis for analyzing and evaluating numerous more complex aspects of neuronal morphology.

In the following section, each of the parameters will be defined and the techniques utilized to obtain them described. Then the distribution of these parameters in several cell types will be presented.

PARAMETERS OF NEURITE SHAPE From an analysis of the length, diameter, and spatial orientation of the dendritic trees (see Figures 1 and 2), the following fundamental parameters have been abstracted:

1. stem diameter (D),
2. terminal-segment diameter (T),
3. segment taper (ΔA),
4. segment length (L),
5. branch power (n),
6. ratio between cross-sectional areas of daughter branches (R),
7. spatial orientation of segments (including branch angle).

FUNDAMENTAL PARAMETERS OF NEURITE SHAPE

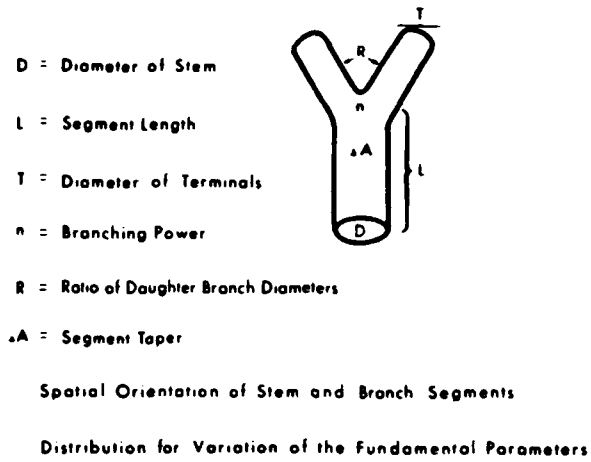


FIGURE 2 Neuronal shapes can be defined by seven variable parameters that describe the form of neuronal arborizations. Variations in the distribution of these fundamental parameters over the tree produce differences in tree types as well as the individual characteristics of neurites.

These parameters describe dendritic trees in the following way. The base of the tree, the *stem segment* (with diameter D), is divided at the first branch point. Here the stem length is determined and two daughter branches formed. (Trifurcation sites occur, but so infrequently that they will not be considered here; see Berry and Bradley, 1976a.) At each successive branch site, *segment lengths* (L) are determined, and the diameter is reduced until the *terminal diameter* (T) is reached. The division of the cross-sectional area of a parent segment into daughter branches is described by two parameters, the *branch power* (n) and the *daughter-branch ratio* (R). Branch power relates the cross-sectional area of the parent segment to the total cross-sectional area of the daughter segments. The daughter-branch ratio is the ratio of the cross-sectional area of the two daughter branches. Between branch points, the diameter may decrease as a *segment taper* (ΔA). Also at branch points the three-dimensional *orientation* of the segments is defined by the direction of the segments with respect to the soma.

The size of the dendritic tree is governed by the stem and terminal diameters, branch power, segment taper, and segment lengths; shape is defined by all seven parameters. Variations in these parameters contribute to the variety and complexity of neuronal arborizations.

THREE-DIMENSIONAL RECONSTRUCTION AND PARAMETER EXTRACTION Three major approaches are cur-

rently in use for tracking neurites in dye-injected or Golgi-impregnated neurons. The first, designed by Glaser and Van der Loos (1965) and adopted by Wann et al. (1973), employs a fixed cursor located in the center of the microscopic field while the X, Y, and Z coordinates are obtained by operator-controlled stage movements and focusing. The second approach employs a movable cursor in conjunction with stage displacements. This method facilitates the manual tracking operations (Llinás and Hillman, 1975; Hillman, Llinás, and Chujo, 1977; Paldino and Harth, 1977a; Lindsay, 1977b) and was used to develop a semiautomatic method (Garvey et al., 1973; Coleman, West, and Wyss, 1973; Coleman et al., 1977). The last approach records the structures by interactive or automatic extraction of the perimeters from profiles of neurons and their processes in serially sectioned preparations (Levinthal and Ware, 1972; Reddy et al., 1973; Selverston, 1973; Levinthal, Macagno, and Tountas, 1974; Glasser et al., 1977; Hillman, Llinás, and Chujo, 1977). This modern version of the wax sheet reconstruction method (Born, 1883) not only allows viewing from every perspective (Hillman, Llinás, and Chujo, 1977), but also provides the necessary data base for an analysis of global parameters such as surface area, volume, and process lengths (Glasser et al., 1977; Hillman, Llinás, and Chujo, 1977).

In our approach, the Cartesian coordinates and diameters of neuronal somata and their processes were recorded with the aid of a graphics computer interfaced to a light microscope (Llinás and Hillman, 1975; Hillman, Llinás, and Chujo, 1977). The system has been specifically configured to provide maximum resolution, 0.2–0.4 μm, in X and Y. Data points are defined for the origin of the dendrites (at the soma), changes in the course of the process, branching sites, terminals, and cut ends (for alignment to subsequent sections).

The data consist of the soma diameter and the length, diameter, and spatial orientation of each dendritic segment. Analysis programs extract and compare values from the data file for parameters of different cells and cell types and compute values for the global parameters of arborizations (e.g., total length, volume, and surface area). Finally, this information is displayed as bar graphs or point plots. The examples included in this paper were obtained from an analysis of rat and cat cerebral pyramidal cells, cerebellar Purkinje, stellate, and granular cells, and motoneurons (see Figure 1). Analysis at the light-microscopic level was based on Golgi material, while subcellular structures were studied from electron micrographs. In ultrastructural studies, the entire regions of the dendritic fields of Purkinje cells, motoneurons, and pyramidal cells were analyzed for microtubules and neurofilaments. These records were quantitated on an image-analysis computer and recorded with reference to the diameter of the respective dendritic profile.

The Fundamental Shape Parameters

Stem diameter (D) The diameters of stem dendrites range from over 10 μm to less than 1 μm. This range is largely due to the variable diameters of individual primary dendrites in multipolar neurons (Bok 1936a,b; Haggar and Barr, 1950; Chu, 1954; Bok, 1959; Balthasar, 1962). The sum of these diameters shows little variability (Table I). Furthermore, for a number of multipolar cell types, this summed cross-sectional area was found to bear a consistent relationship to soma size (Figure 3). This is in agreement with Rall (1959), who found that motoneuron stem diameters (represented by the sum of their diameters to the 3/2 power "combined trunk parameter") correlated well with soma surface area. The significance of this finding lies in the relationship between the summed stem cross-sectional area and the volume of the dendritic tree (see below). The diameters of cells with only one stem dendrite are less variable (e.g., see Purkinje cells, Table I); although this parameter has a similar correlation with soma size, the dendritic cross-sectional area is somewhat smaller.

Terminal diameter (T) As illustrated in Figure 4, the minimum diameter of terminal segments is sharply delineated (there are very few segments smaller than 0.5 μm). Furthermore, although terminal diameter is the least variable parameter (Table I), two groups are found, the first with a mean diameter of 1.1 μm, the second with a mean of 0.76 μm.

Segment taper (ΔA) In a number of cell types, the

Figure 3 The sum of cross-sectional areas of all stem dendrites that arise from the soma shows a correlation to the soma diameter. Granule cells, pyramidal cells, and motoneurons form a series of somas with increasing diameter, and when plotted as the square with the cross-sectional area for the sum of emerging dendrites, a strict correlation is found. The slope of this relationship (a) is 3/2, indicating that this dendritic cross-sectional area is proportional to a spherical volume related to the soma. A second parallel slope (b) is formed by Purkinje cells. Thus multipolar neurons have more cross-sectional area emerging from a soma of particular size than do unipolar cells of comparable size. These data are consistent with the relationship of the "combined trunk parameter" to soma surface area (Rall, 1959). Soma and stem diameters obtained by computerized interactive recording of Golgi preparations. (Unpublished results.)

TABLE I

Parameters of form

Cell type	Cross-sectional area (μm^2)		Diameter of terminal segments[a] (μm)	Length (μm)		Branch power	Taper	Daughter-branch ratio
	Individual stem dendrites	All stem dendrites		Terminal segments	All segments			
Pyramidal	12.8 ± 13 (100)[a]	76.8 ± 21 (26)	1.17 ± 0.34 (29)	120 ± 59 (49)	70.8 ± 65 (91)			
Apical dendrites						1.99 ± 0.79 (40)	marked[b]	2-6
Basal dendrites			1.11 ± 0.33 (29)			2.28 ± 0.89 (39)	not evident	<2
Purkinje	109.3 ± 14 (13)	109 ± 14 (13)	1.04 ± 0.30 (29)	15.2 ± 10 (79)	11.6 ± 9.2 (79)	2.36 ± 1.2 (51)	none (except stem)	low and high
Granule	1.51 ± 0.79 (52)	5.48 ± 0.94 (17)	0.66 ± 0.25 (38) 0.76 ± 0.29 (32)	4.58 ± 3.7 (81)	10.7 ± 8.4 (78)	2.58 ± 1.8 (71)	not evident	low
Stellate	2.75 ± 1.35 (49)	11.0 ± 3.4 (31)	0.73 ± 0.30 (41)	23.9 ± 29 (100)[a]	31.7 ± 23 (72)	2.24 ± 1.2 (53)	not evident	low
Motoneuron	88.4 ± 56 (64)	886 ± 71 (11)	—	—	—	1.69 ± 0.48 (28)	marked[c]	low

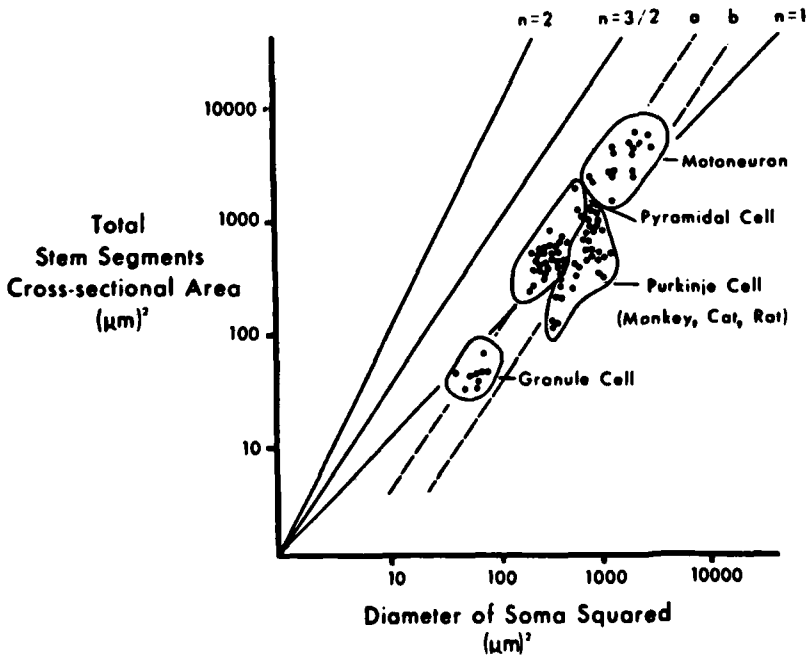
Mean values ± S.D. Numbers in parentheses are coefficients of variation, defined as the standard deviation as a percentage of the arithmetic mean.

[a] As measured near the final branch point.

[a] Test assumes a normal distribution and is actually inappropriate for these skewed distributions.

[b] Unpublished observations.

[c] See Barrett and Crill (1974a) and Lux, Schubert, and Kreutzberg (1970).



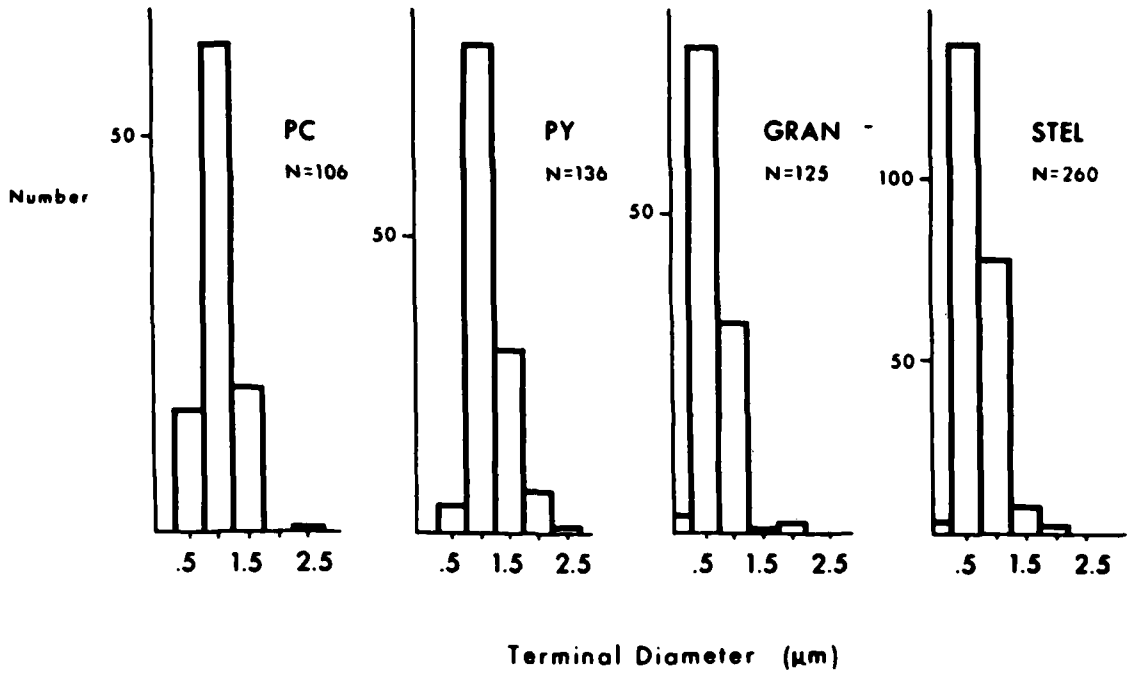


FIGURE 4 Diameters of terminal segments are consistently constrained by minimal and maximal limits. Purkinje (PC) and pyramidal (PY) cells tend to have somewhat larger terminal diameters than do granule (GRAN) and stellate (STEL) cells. These differences may be related to the fact that Purkinje and pyramidal cells have numerous spines along their course while granule and stellate cells for the most part lack spinelike structures. The diameter was recorded on Golgi-impregnated neurons near the final bifurcation. (Hillman, Llinás, and Iberall, unpublished results.)

diameter of the dendritic process decreases between branch points. Taper varies between cell types; a marked taper has been found in some cells while others show none (see Table I; Lux, Schubert, and Kreutzberg, 1970; Barrett and Crill, 1974a).

Segment length (L) Marked variations are found in the length of dendritic segments (Table I; Figure 5; see also Smit, Uylings, and Vledmaat-Wansink, 1972; Berry and Bradley, 1976b; Lindsay, 1977c). The shortest segments result from trichotomous branching in which the daughter branches arise very close together (Berry and Bradley, 1976a). In contrast, terminal segments can reach from 30 μm to over 200 μm (Bok, 1936a; Peters and Bademan, 1963). Intermediate and stem segments have a smaller group range. Pyramidal cells are unusual in that their terminal-segment length is drastically different from that of other segments.

Since the distribution of segment lengths over the tree varies widely, this factor no doubt contributes to the range of shapes as well as sizes of dendritic trees.

Branch power (n) Tree shape is influenced by changes in the cross-sectional area at branch points. One way to assess if such a change has occurred is to determine the branch power n, defined as the exponent that satisfies

$$D^n = \sum_j d_j^n , \qquad (1)$$

where D is the diameter of the parent dendrite and d_j is the diameter of the jth daughter dendrite (Figure 6). This equation describes a geometric relationship between tree branches and has been found to be of great importance in describing the cable properties of neurons when both *geometric* and *electric* factors must be brought together (Rall, 1959).

A first step in developing the equation for branch power is to determine the input conductance of an individual dendritic segment. When considering elec-

FIGURE 6 Two branch power rules: $n = 3/2$ and $n = 2$. The 3/2 power law, postulated by Rall (1959), was believed to be important for conservation of electrical charge for electronic spread from distal parts of the dendritic tree. In contrast, a conservation of cross-sectional area results from a square law when the volume remains constant throughout the dendritic tree. The equivalent volume and equivalent electrical cylinders are illustrated for each rule. Surface area (not shown) would be considerably increased.

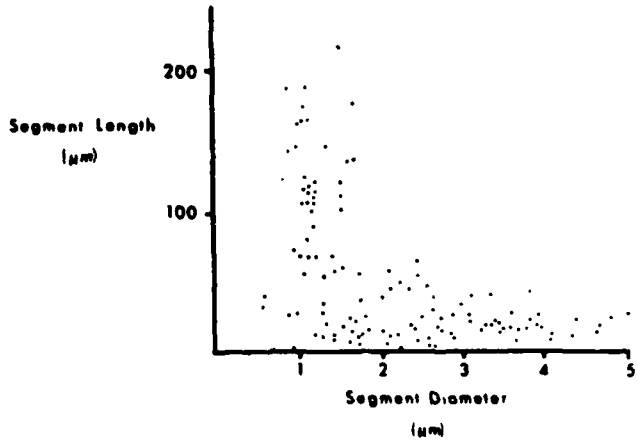


FIGURE 5 Segment lengths are variable within dendritic trees; however, there are characteristic limitations on the maximum lengths. Here in an example of pyramidal-cell dendrites, stem segments and intermediate segments extend up to 50–60 μm (dots) while terminal segments (squares) exceed 200 μm. Note the constrained diameter range of the terminal segments. The diameters were recorded near the final branch point, and the terminal segments were defined by the Strahler ordering method.

trotonic properties of neurons, it is common practice to idealize axons and, therefore, dendritic segments as uniform cylinders.

The input conductance G of a cylinder is proportional to the 3/2 power of the diameter d:

$$G \propto (R_m R_i)^{-1/2} d^{3/2}, \tag{2}$$

where R_m is the resistance across a unit area of membrane and R_i is the specific resistance of the internal medium.

In order to facilitate the treatment of electrotonus in dendrites, the tree was collapsed into a uniform cylinder. In order to perform this simplification, however, passive electrical properties of cables dictate that the input conductances on each side of a branch point must be the same, so that charge is not lost. Thus

$$G_{\text{parent}} = \Sigma\, G_{\text{daughters}}. \tag{3}$$

Combining Equations 2 and 3,

$$(R_m R_i)^{-1/2} D^{3/2} = (R_m R_i)^{-1/2} \sum_j d_j^{3/2}, \tag{4}$$

so that

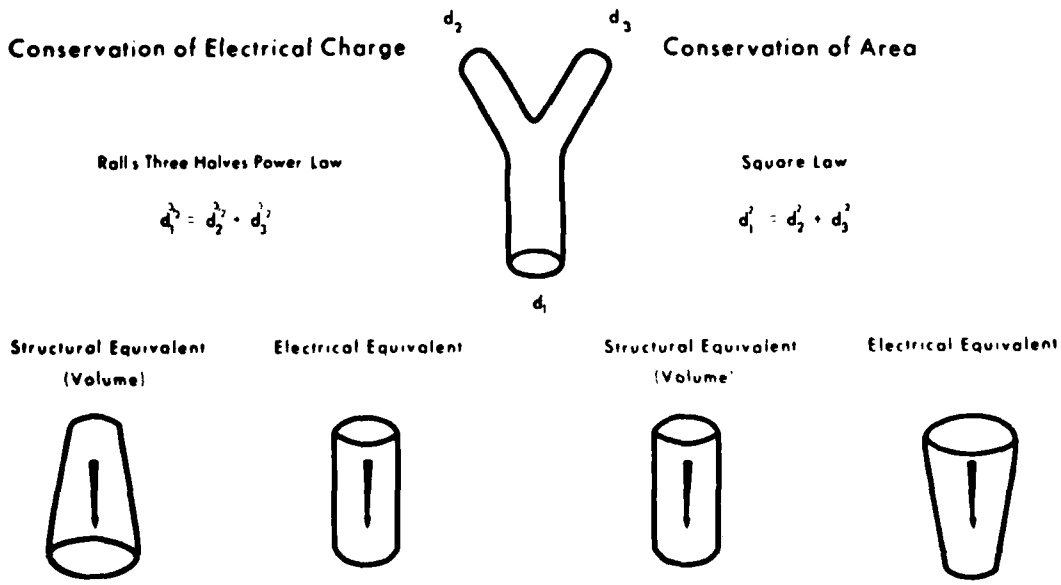
$$D^{3/2} = \sum_j d_j^{3/2}. \tag{5}$$

This has come to be known as the 3/2 power rule.

On the basis of published photographs of Golgi-impregnated motoneurons, Rall (1959) found that the 3/2 power rule, derived from theoretical consideration, did indeed hold for this cell type. More recently, Lux, Schubert, and Kreutzberg (1970), Barrett and Crill (1974a,b), and our own studies (Figure 7) have verified these first measurements.

Similar measurements for branching in pyramidal and Purkinje cells do not support the 3/2 power rule

Branch Power Rules



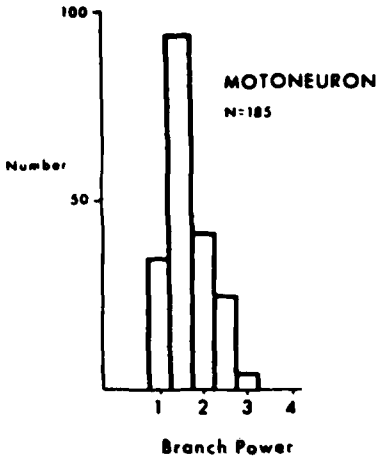


FIGURE 7 Histogram for branch power in motoneurons. The values of branch power for individual motoneuron branches cluster near 3/2. Branch power was obtained by solving the branch power equation by successive approximation (Newton-Raphson algorithm). The spread for the smaller population of the branch powers is primarily due to the sensitivity of the method for very large daughter-branch ratios. Data recorded from Golgi preparations of rat spinal cord.

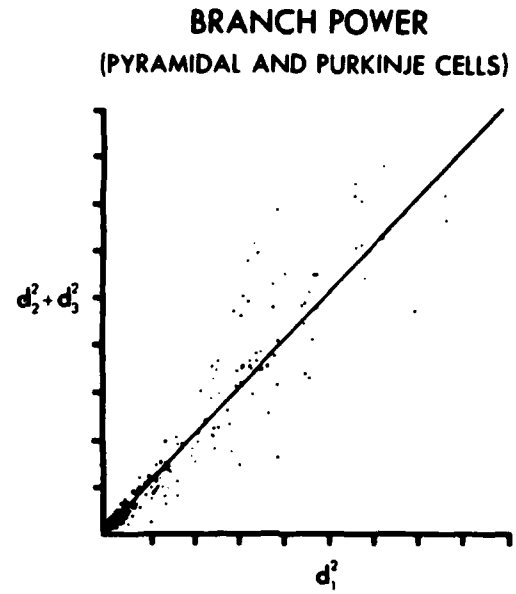


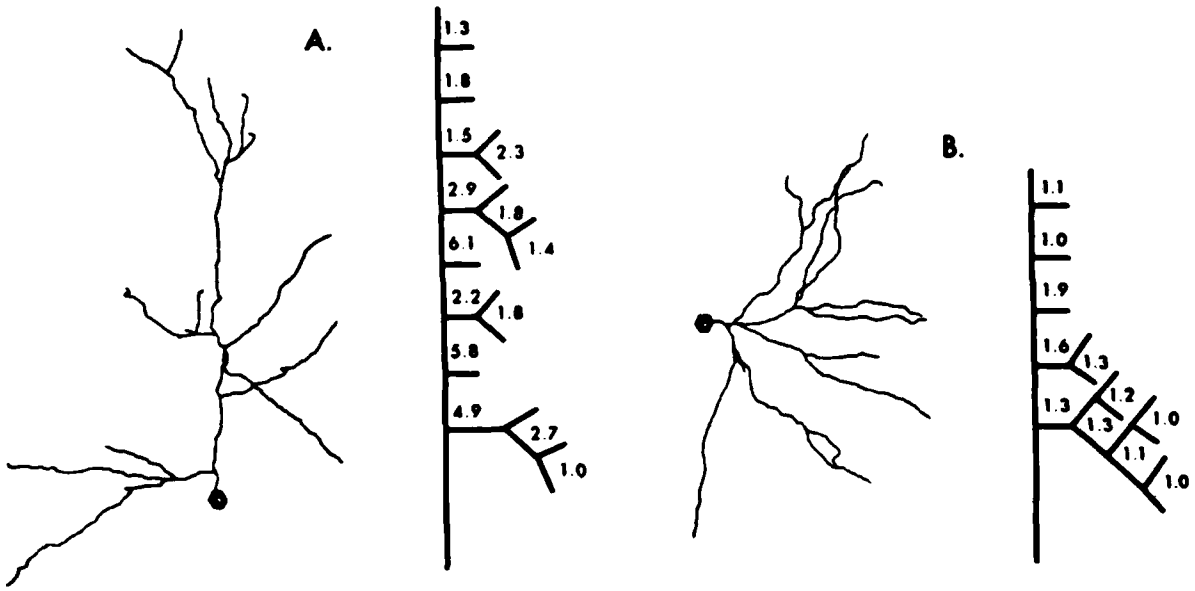
FIGURE 8 In pyramidal and Purkinje cells, branch power was determined by a graphic method for solving the power equation. Plots were made for a 3/2 power and for a power of 2. It was found that, on average, a least-squares fit, for a power of 2, had a slope of one (this illustration). Thus $d_1^2 = d_2^2 + d_3^2$, where d_1 is the diameter of a parent and d_2 and d_3 are the diameters of its daughter branches. (Rat pyramidal and Purkinje cells.)

(Table 1; Figure 8), but rather approximate a square rule. In these cells charge is not conserved across branch points. The important point to be considered here, however, is that in cells following the 3/2 power rule, cross-sectional area is reduced at branch points, but in those following the square rule, cross-sectional area is conserved. The functional significance of these power rules is beyond the scope of this paper. (For a detailed discussion see Jack, this volume.)

Daughter-branch ratio (R) The ratio of the diameters of daughter-branch processes at each bifurcation defines the daughter-branch ratio R (the usual range is 1–10). Equal daughter branch diameters ($R = 1$) are seldom observed; values of 1.5–4 are more often found. Clearly this parameter (together with branch power) alters tree shape at branch points. For example, at a bifurcation having a daughter ratio of 4, the larger-diameter daughter branch will require many further bifurcations to reach the terminal diameter than will the smaller daughter branch. If this disparity is common, the entire pattern of branch distribution is affected. From this relationship alone it is evident that daughter-branch ratio can have a deep influence on tree shape. Specifically, pyramidal-cell apical dendrites have very high (2–6) daughter-branch ratios and therefore have numerous unequal bifurcations. This produces a relatively narrow tree growing from a single thick trunk. In contrast, the basal dendrites have a daughter ratio less than two, and the branching number at each level is almost constant from the stem to the terminal segments (Figure 9B). Thus a "bushy" tree is produced (compare Figures 1A, B and 9).

Although mammalian Purkinje cells also have close to equal bifurcations throughout the tree, numerous smaller daughter branches are also found on large dendrites. This interspersion of small-diameter branches ($R > 2$) results in a very dense arbor.

Spatial orientation A complete description of dendritic trees requires defining the tridimensional orientation of each segment. Of the many schemes developed toward this end, two are particularly useful: the application of principal-component analysis (Brown, 1977) and the use of polar coordinates (Paldino and Harth, 1977b). Other methods in use only approximate or describe limited aspects of the three-dimensional relationship of processes (see, for example, Uylings and Smit, 1975; Lindsay, 1977d). We have instituted a simplified spherical coordinate system very similar to that used by Paldino and Harth (1977b). This system is capable of defining spatial relationships with sufficient accuracy to catalog the position of individual segments and is at the same time flexible enough to be used to define the position



Apical Dendrite **Basal Dendrite**

FIGURE 9 Daughter-branch ratios vary consistently over various tree types. Topological patterns are displayed according to Berry et al. (1975) for each reconstructed dendrite. The numbers at the bifurcations are daughter-branch ratios. In apical dendrites of pyramidal cells (A) the bifurcations of the main stem have very high daughter-branch ratios (2–6). Basal dendrites (B), on the other hand, have daughter-branch ratios less than two. We believe that this factor is most important in determining topological type in neuronal form.

"Topological type" refers to branching patterns irrespective of length, diameter, and orientation. According to the dendritic model this relationship in neurites can be modified by four fundamental parameters: branch power, segment taper, terminal diameter, and daughter-branch ratio. The most marked influence is that of the daughter-branch ratio; the other three parameters are relatively consistent within cell types and thus do not significantly alter the number of branch levels necessary to reach the terminal diameter.

of an entire neuron including its dendrites, soma, and axon. Thus the location and extent of a great number of neurons can be compiled for comparative analysis. (This method is discussed below.)

GLOBAL PARAMETERS The combination of certain of these fundamental parameters (e.g., segment length and diameter) forms a new set, that of global parameters. These parameters are important because they go beyond specific aspects of shape to provide information about the size and orientation of trees. It is at the level of global parameters that cell form is most clearly related to function.

Volume Volume is a measure of the size of the dendritic tree and has a close correlation to soma size (Mannen, 1966). In fact, all measurable global parameters (surface area, summed segment length, maximum continuous length) are correlated with soma size (Sholl, 1955; Mannen, 1966; Gelfan, Kara, and Ruchkin, 1970). Likewise, we have found that the volume of each individual dendritic tree is correlated with the diameter of its stem (Figure 10B). Thus the distribution of each of these parameters among individual dendrites can be discerned by measuring the diameters of the stem dendrites. The correlation of global parameters with soma size is to be expected as all these parameters ultimately reflect the amount of cytoplasm and membrane in the dendrites. These are composed of molecular entities which must be maintained by the machinery of the cell factories located, for the most part, in the cell soma (Llinás and Iberall, 1977).

Surface area The dendritic surface area is the largest interactive part of the cell, comprising more than 80% of the surface in some neurons (Mannen, 1966). Furthermore, dendrites have recently been recognized as having presynaptic (Shepherd, this volume) as well as postsynaptic and nonsynaptic (Kreutzberg, this volume) interactions. The surface area then reflects the capacity of a cell to receive (Rall, 1970) and, in many cells, to contribute inputs.

Length The length of dendritic processes has been

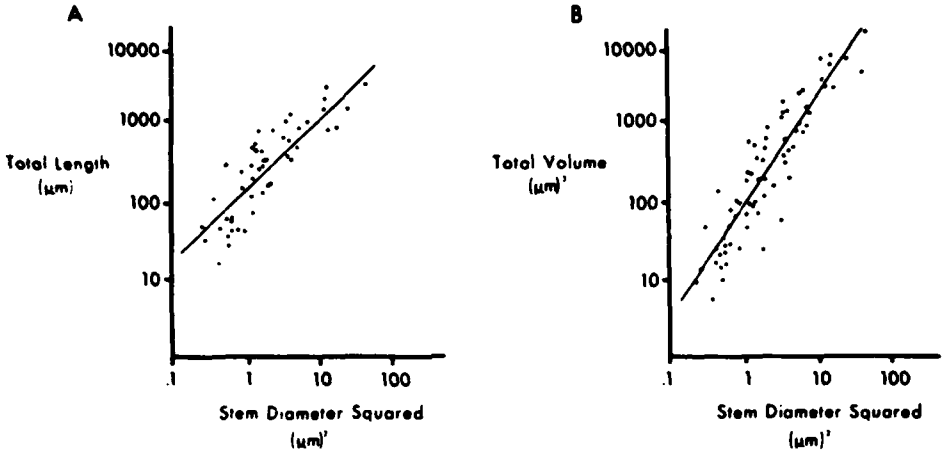


FIGURE 10 Correlation for stem diameter with total length and volume of individual dendritic trees. Each point represents a single tree from a granule, stellate, or pyramidal cell. (Unpublished results.)

measured in a number of ways (Sholl, 1953; Fox and Barnard, 1957; Schadé and Van Groenigen, 1961; Glaser and Van der Loos, 1965; Mannen, 1966; Gelfan, Kara, and Ruchkin, 1970; Wann et al., 1973; Hillman, Chujo, and Llinás, 1974; Lindsay and Scheibel, 1974). One of the most useful measurements is the maximum continuous dendritic length (Gelfan, Kara, and Ruchkin, 1970) since, when this value is combined with the process direction (with reference to the soma), the vector quantity *spread* is formed. Information is then available concerning how far and in which direction a cell sends its dendrites. Used in this way, length is the only global parameter to combine size with another aspect of neuronal morphology. Furthermore, when length and diameter are combined with orientation (see below), a complete description of the shape and size of a cell emerges.

Orientation Orientation, as a fundamental parameter (see above), describes the location of dendritic segments and processes with respect to each other and to the soma. As a global parameter, orientation first places the neuron with reference to major brain landmarks (e.g., hippocampus, corpus callosum) and then in the context of neighboring structures (e.g., other neurons, glia, blood vessels). From a functional viewpoint, the most important of these structures are the neighboring neurons, for these comprise the circuit in which the cell functions.

There is an extensive literature demonstrating the specificity of cell orientation in both laminar and nuclear structures (e.g., Ramón y Cajal, 1909, 1911; Lorente de Nó, 1934; Palay and Chan-Palay, 1974). Other studies have been most valuable in determining precise shifts in dendritic projections following perturbations (Clendinnen and Eayrs, 1961; Schadé and Van Groenigen, 1961; Peters and Bademan, 1963; Colonnier, 1964; Mungai, 1967; Wong, 1967; Schadé and Caveness, 1968; Smit, Uylings, and Vledmaat-Wansink, 1972; Smit and Uylings, 1975; Coleman et al., 1977; Lindsay, 1977d; Paldino and Harth, 1977b). Despite the important contributions of this recent work, a universal method capable of defining spatial relationships with sufficient accuracy to catalog and compare neurons and their locations has not been developed. We have instituted a simplified spherical coordinate system, similar to that recently used by Paldino and Harth (1977b), which may meet this need (cf. Peterson, 1955).

According to this scheme, the primary reference point is the soma center, around which an imaginary sphere is drawn. The intersection of a dendritic process with the surface of the sphere is defined by latitudinal (range 0 to 180 degrees) and longitudinal (range 0 to 360 degrees) coordinates. The application of this system to a cortical pyramidal cell is illustrated in Figure 11. In this example latitudinal coordinates greater than 90° indicate that the dendrite projects away from the pial-glial surface (Figure 11B). Longitudinal coordinates less than 180° indicate that the processes project caudally.

Essentially this method utilizes two common references to yield two coordinate points for each site of interest. The orientation of dendritic stem segments and segments at branch points can be defined, thus facilitating descriptive comparisons and cataloging of spatial orientation. (The cell body itself is located on a three-dimensional Cartesian coordinate system which is fixed for a given brain or region.)

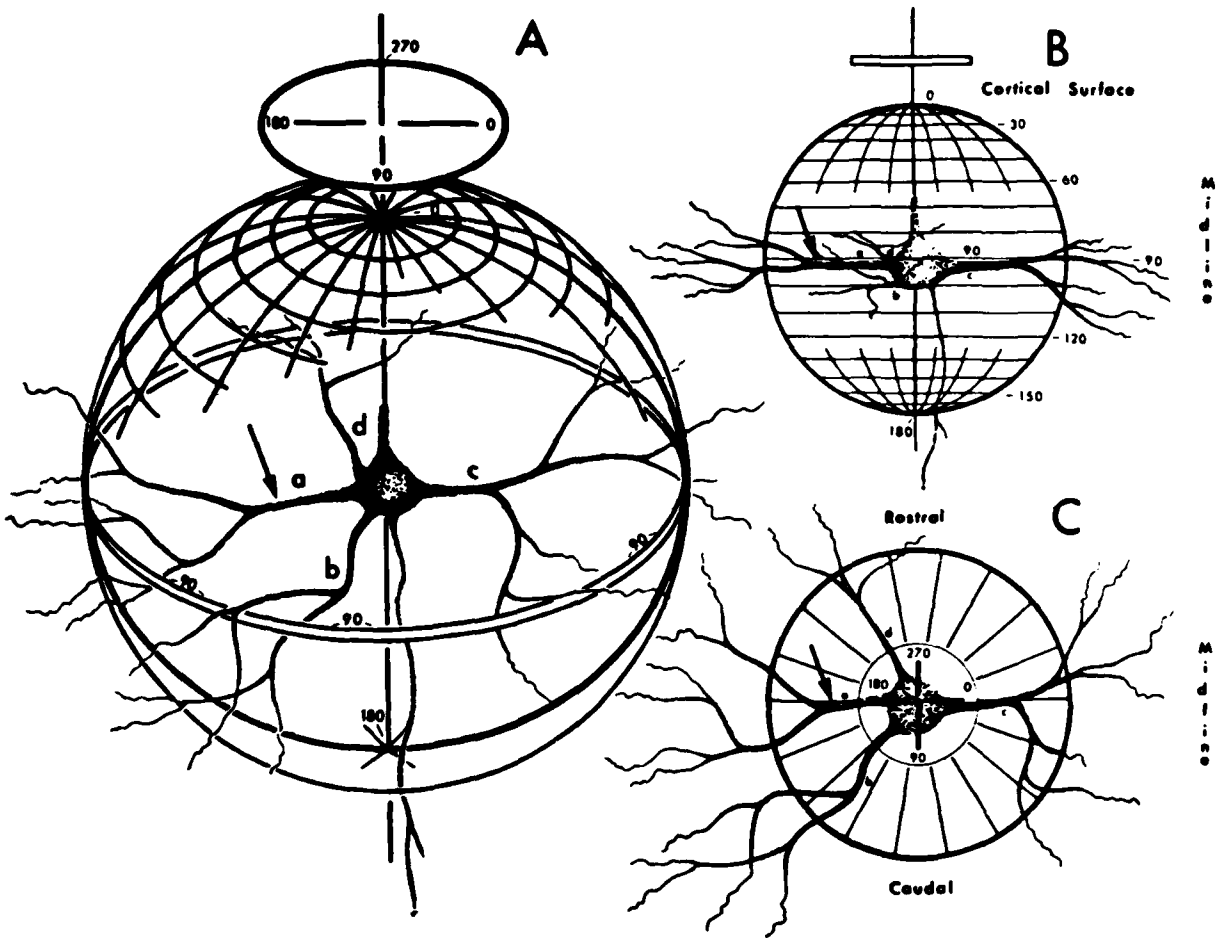


FIGURE 11 A spherical-coordinate method for defining orientation and comparing dendritic branching. Two reference axes are established in order to limit the number of coordinates to two for each segment. One reference is an axis that might, for example, extend from a surface of a laminar structure perpendicularly into the depth of the nervous tissue. This axis passes through the center of the soma, tree base, or branch point and represents the center of the sphere (A). The orientation of the dendritic segments relative to this axis is defined by two coordinates: First, a latitudinal (La) coordinate extends from 0 to 180°, representing a north-south relationship (B), where north is the surface of a laminar structure in the brain. A dendrite that extends above the equatorial zone, 90°, projects toward the surface and has an (La) angle within the 0–90° range. Those below the equatorial zone have angles between 90° and 180°. Second, the orientation of these dendrites around this north-south axis are made by longitudinal (Lo) coordinates that range from 0° to 360° (C). The zero reference for the longitudinal coordinates is related to another common landmark such as an afferent, efferent, or identifiable dendrite that is consistently limited to the same side of this north-south axis. It is important that this reference reflect the convolution and the curvature of the cortex so that accurate comparisons can be made between the cells.

Topological types Each dendritic tree can be categorized according to its topological type (Berry et al., 1975). The assignment of a tree to a particular category depends on the patterning of its segments and is independent of their length, diameter, or orientation (Hoopen and Reuver, 1970). The four fundamental parameters determining the number of branch points (stem diameter, terminal diameter, branch power, daughter-branch ratio) influence these patterns. The most variable of these (next to stem diameter), daughter-branch ratio, makes the most significant contribution and plays a major role in determining the topological type of a dendritic tree.

This approach to categorizing neuronal trees has been used to compare the probability for the occurrence of particular branching patterns in normal adults (Berry et al., 1975) with the probability of their occurrence in animals at various stages of develop-

ment (Berry and Bradley, 1976b) or following alterations to the adult nervous system (McConnell and Berry, 1978).

Arborizational taper Taper in neuronal branches is extremely important because of its significance in electrotonus (Rall, 1959, 1962). A means of evaluating this factor is to determine the power relationships for each arborization. This parameter expresses both branch power and segment taper as a power for the entire dendritic tree and is given by the relationship

$$D_s^n = d_{t_1}^n + d_{t_2}^n + \cdots + d_{t_N}^n = \sum_j d_{t_j}^n,$$

where D_s is the diameter of the stem and d_{t_j} is the diameter of the jth tree terminal. The approach is a modification of Rall's (1959, 1962) branch-power equation and is a means of approximating changes in cross-sectional area for the entire tree. The solution of the power equation is generated by software for a method of successive approximation. Furthermore, the difference in cross-sectional area between the stem and the sum of the terminals can be combined with length to establish tree taper. Hillman and Gelbfish (unpublished observations) have determined the arborizational taper of basal dendrites of pyramidal cells and produced encouraging results. Further work will determine its usefulness.

Structural components of neuronal form

Having established basic parameters of neuronal shape, we can ask what intra- or extracellular elements provide the structural basis for neuronal form. The most general hypothesis is that of Porter and Tilney (1965), who suggested that the subcellular elements form a structural core underlying the shape of cells. Basically it is assumed that cell shapes deviating from a sphere require an intracellular scaffold or framework. Thus elongation of cells would require the support of a cytoskeleton composed of specific subcellular organelles (Tilney and Porter, 1967; Yamada, Spooner, and Wessells, 1970). Following the ultrastructural identification of microtubules and neurofilaments by Schmitt and Geren (1950), investigators have suggested that these elements may be the principal components of this cytoskeleton.

SUBCELLULAR STRUCTURES Quantitative analysis of neuronal ultrastructure has been limited, for the most part, to the soma and axon. Studies determining soma size and the nucleus-to-cytoplasm ratio are the most numerous (Bok, 1936b, 1959; Sholl, 1955; Gelfan, Kara, and Ruchkin, 1970). Little information has been forthcoming regarding subcellular structures that may be correlated to nuclear or somatic size.

In axons, microtubules and neurofilaments have been the primary focus for analysis (see Schmitt, 1968; Wuerker and Kirkpatrick, 1972). Most quantitative studies of these structures have been limited to correlating tubule and filament number to axon diameter (Friede and Samorajski, 1970; Zenker, Mayr, and Gruber, 1973). However, there have been two attempts to determine whether axonal tubules branch or are lost, after they leave the soma, by comparing the number of tubules in the axon after it emerges from the central nervous system with the number in the telodendron (Weiss and Mayr, 1971; Zenker and Hohberg, 1973). These studies are not in agreement, and further investigation is needed.

Qualitative descriptions of filaments and tubules in dendrites of pyramidal and Purkinje cells (Wuerker and Kirkpatrick, 1972), cerebral cortical cells (Peters, 1968, 1971), motoneurons (Wuerker and Palay, 1969), and Clark's column cells (Smith, 1973) have found that when dendrites are viewed in cross section, tubules are evenly dispersed but filaments are usually found in groups (Figure 12). Also, the densities of these elements vary among cell types.

In studies counting both filaments and tubules, we have verified these findings for three cell types (Figure 13A, C; Table II). Specifically, pyramidal- and Purkinje-cell dendrites maintain a constant microtubular density (Figure 13A), whereas a second pattern is found in motoneurons. Here the tubule density decreases sharply toward the base of the tree while the smallest branches have concentrations close to those found in pyramidal-cell dendrites of similar diameter (Figure 13C).

The distribution of neurofilaments also depends on cell type. Motoneurons and pyramidal cells have a relatively constant filament density. In fact, in most motoneuron processes, filament density surpasses tubule density (see Figure 13A, B). Purkinje cells, for the most part, lack neurofilaments.

TABLE II
Density of cytoskeletal structures

Cell type	Microtubules (μm^2)	Neurofilaments (μm^2)	T/F ratio
Pyramidal	70.8	4.99	14:1
Purkinje	28.4	1	–
Motoneuron		38.6	
base	31.9		1:1
terminals	100		2.5:1

T, microtubules; F, neurofilaments.

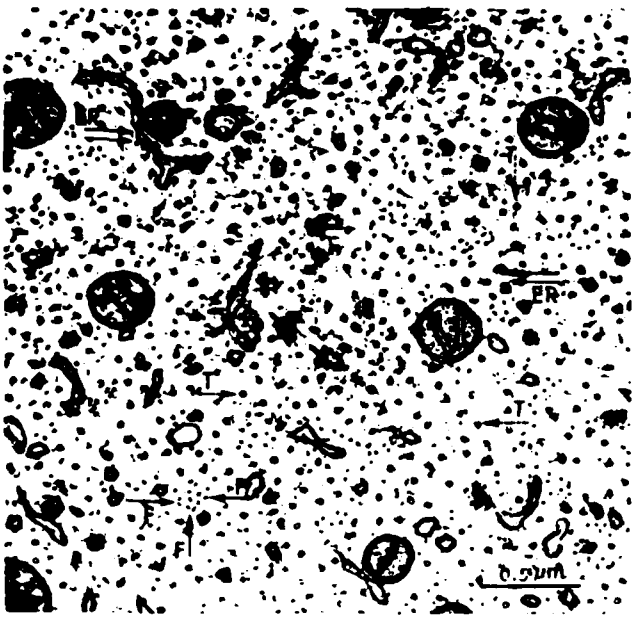


FIGURE 12 Distribution of tubules and filaments in a motoneuron dendrite. Here an electron micrograph from the ventral horn of the spinal cord demonstrates the general distribution of microtubules and neurofilaments in cross section. Preparation stained with uranyl acetate and lead citrate. (T, tubules; F, filaments; FR, smooth endoplasmic reticulum.)

CORRELATIONS BETWEEN SHAPE PARAMETERS AND SUBCELLULAR COMPONENTS One of the major questions generated by these findings is whether microtubules and neurofilaments, or indeed other cellular organelles, represent components of a structural framework that helps define the shape of neuronal somata and processes. This is a problem we are just beginning to explore, and the following results are the first from an approach that promises to play an important part in our study of neuronal form.

Stem diameter The clearest correlation between cytoskeletal structures and parameters of form is found with stem diameter. In fact, this is true for the diameter not just of the stem, but of all segments. For this reason all segments, including stem and terminal segments, will be considered here. There is a clear correlation between number of microtubules and neurofilaments and segment diameter in Purkinje cells (Figure 12), pyramidal cells (Figure 13A), and motoneurons (Figure 13C). In order to understand the relationship of microtubules and neurofilaments to segment diameter, one must consider not only the increase in the *number* of these elements with increases in diameter, but also (1) changes in the *density* of these elements and (2) the *ratio* of microtubule to neurofilament density (Table II). The three cell types studied will be considered sequentially, beginning with the simplest case—the Purkinje cell.

In Purkinje cells, microtubule numbers increase with segment diameter and maintain a constant, although relatively low, density ($28/\mu m^2$). This cell contains almost no neurofilaments; therefore the tubules are clearly correlated with and may play a major role in establishing segment diameter (Figure 13B). Pyramidal cells have a higher density of tubules ($71/\mu m^2$) and also contain neurofilaments. Although the tubule-to-filament ratio is 14:1, the filament density is fairly constant, and both these elements doubtless contribute to determining the segment diameter in this cell type (Figure 13B).

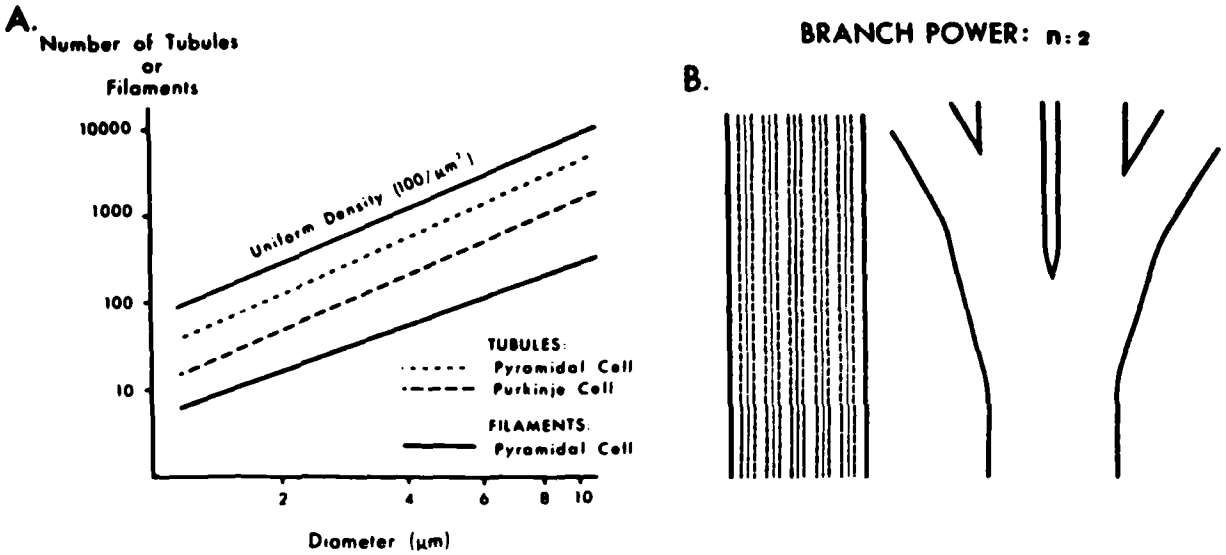
The most complicated case is the motoneuron. Here tubule density decreases with increased segment diameter, but there is a significant number of filaments and their density is constant (Figure 13D). Again, at least one element, the neurofilament, can be clearly correlated with diameter.

This correlation between segment diameter and cytoskeletal elements suggests that these structures may provide the framework for dendritic processes and so may contribute significantly to determining the diameter of the branches of dendritic trees. While the observation that the total stem cross-sectional area is proportional to soma volume is intriguing (see Figure 3), very little can be said about this relationship since little is known about the ultrastructural basis of soma size.

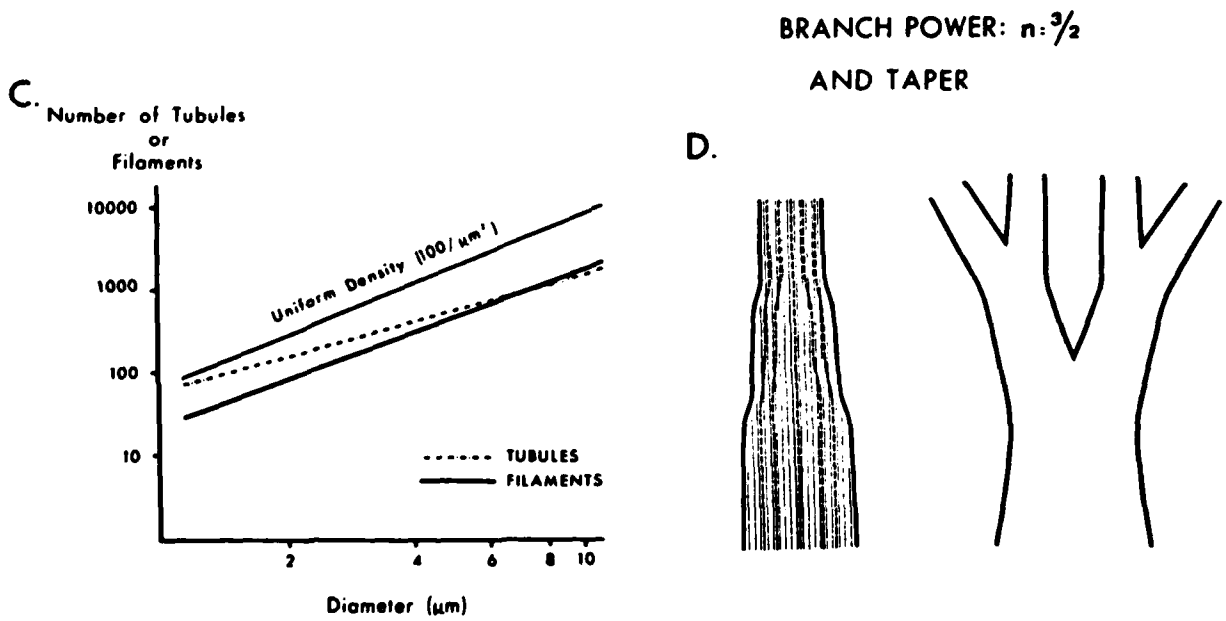
Terminal diameter There are two aspects of terminal-segment diameter that bear further consideration: (1) the sharp lower size limit, and (2) the difference between terminal segments with spiny branches and those without. First, the clear and consistent relationship between segment diameter and tubule number suggests that the smallest diameters may, in fact, reflect the space needed to contain a certain number of tubules—the number required to maintain the process. This suggestion finds support in the observation that terminals bearing spines are wider. Preliminary studies of electron micrographs indicate that these processes do not contain more tubules but have more endoplasmic reticulum than do the narrower terminal processes.

Segment taper Although the tapering between branch points decreases the process surface area, it reduces the process volume even more. Thus, as a segment tapers either the *density* of the cytoskeletal elements must increase or their *number* must decrease. These results suggest that the latter may be the case and that tapering is associated with a decrease in neurofilament number. Although this re-

PURKINJE AND PYRAMIDAL CELL DENDRITES



MOTONEURON DENDRITES



quires further investigation. I have found that while cells with the most marked taper—motoneurons—contain the highest density of filaments ($38.6/\mu m^2$), this density remains constant as the diameter of processes decreases.

The ultrastructural basis for the unusually large taper seen in motoneuron (Barrett and Crill, 1974a) and pyramidal-cell stem segments is different. In these instances the stem taper is correlated with the extension of Golgi complex and rough endoplasmic reticulum (see Peters, Palay, and Webster, 1976) into the base of the arbor.

Branch power The finding of cell types obeying the 3/2 power rule (motoneurons) and the square power rule (pyramidal and Purkinje cells) is especially interesting if one considers the distribution of microtu-

FIGURE 13 Microtubule and filament densities and their correlation to branch power. (A) Tubules (broken lines) within Purkinje- and pyramidal-cell dendritic trees follow a constant density throughout the range of cross-sectional profiles for different parts of the arbor. Pyramidal-cell tubules have the highest density while Purkinje cells are somewhat less concentrated. Filament density (solid lines) in pyramidal cells is nearly uniform. The Purkinje-cell filaments are not prominent or lacking and thus could not be quantitated. (Compiled from data by Hillman, Gelbfish, Llinás, and Iberall.) (B) The diameter of a bifurcating tree (right) and an equivalent cylinder (left) are illustrated for a branch power of 2 (e.g., pyramidal and Purkinje cells). Note that as the segments bifurcate, the diameter decreases while the equivalent cylinder maintains the same diameter over the tree; thus cross-sectional area is conserved. The subcellular filaments (solid lines) and microtubules (broken lines) maintain uniform tubule and filament concentrations. (C) Filaments (solid lines) in motoneurons have a uniform density throughout the extent of cross-sectional profiles for dendrites of different diameters. In contrast, the microtubules in the same profiles decreased in density for profiles with larger diameters. Note that in the larger profiles, the densities of filaments and tubules are nearly equal. It appears that filaments take over the role of maintaining diameter for the base of the tree. The subcellular structures were quantitated in the ventral horn of the rat spinal cord. (Compiled from unpublished data by Hillman and Gelbfish.) (D) The diameter of a bifurcating arbor having a 3/2 power and taper (right) and its equivalent cylinder (left). The reduction in diameter at each bifurcation for branch power 3/2 is significantly greater than that which occurs for branching with a power of two (B above). In the equivalent cylinder each branch point results in a decrease in diameter and thus in cross-sectional area. For the two branches illustrated, the first level represents the diameter for a cross-sectional area from two daughters, while the second represents this equivalent diameter for four daughters. As shown in C, filament density remains constant, thus they end (D, solid lines) or condense into fewer structures for each branch level. Tubules (broken lines), on the other hand, converge centripetally or, conversely, decrease their density toward the base.

bules and neurofilaments, and their relationship to diameter, in these neurons.

Consider first Purkinje and pyramidal cells. Since the cross-sectional area is maintained throughout the arborizations ($n = 2$), the trees can be modeled as simple, uniform cylinders (Figure 13B). (The simplifying assumption that branch power is constant throughout the tree has been made. Our data show a large amount of variability—see Table I and Figures 7 and 8—but this is probably due to the technical difficulties encountered in determining the diameters of small segments.) This finding, coupled with the observation that the microtubules, the major dendritic cytoskeletal elements in these cells, maintain a constant density throughout the tree, supports the hypothesis that microtubules (together with neurofilaments in pyramidal cells) are important determinants not only of process diameter but ultimately of branch power and dendritic volume.

Motoneurons present a more complex case in that neurofilaments and microtubules must both be considered. Because these cells have a branch power of 3/2, the cross-sectional area and volume of each process are reduced at branch points; that is, the dendritic tree must be represented as a tapering cylinder (Figure 13D). (Note that this is a geometric rather than an electrical representation of the cell. Furthermore, the total membrane surface area increases toward the dendritic terminals; thus volume and surface area must be considered separately.) Neurofilaments probably determine this taper, since their numbers must decrease toward the periphery if their density is to be kept constant. The filament number is probably reduced gradually between branch points, and this is reflected in the taper found in the individual segments of these cells (see Lux, Schubert, and Kreutzberg, 1970; Barrett and Crill, 1974a).

In motoneurons the findings that the microtubule density is increased at branch points and that $n = 3/2$ (i.e., the volume decreases at branches) support the view that the number of microtubules is constant in this cell type. This proposal finds support in the work of Weiss and Mayr (1971) who studied microtubule numbers in sensory and motoneuron axons and found that tubule numbers were constant across branch points. They concluded that tubules begin at the soma and extend to the terminals without branching or being lost (this may not be the case in all motoneuron axons; see Zenker and Hohberg, 1973). Hence, our results suggest that tubule numbers are constant in the dendrites of motoneurons and pyramidal and Purkinje cells (in the last two cell types, tubule density and process volume are both constant).

Thus both those cells in which the dendritic volume can be modeled as a uniform cylinder (pyramidal and Purkinje) and those in which it is best modeled as a tapering cylinder (motoneurons) have cytoskeletal elements whose density is constant throughout the dendritic tree: tubules in the former case, filaments in the latter. Cytoskeletal elements can therefore be closely correlated with the diameter of dendritic processes and with their volume. In fact, although there is no direct evidence that these structures "determine" the size of dendrites, I would support this viewpoint rather than its converse (i.e., that the size of the dendrites determines their contents).

Finally, I would propose that the reason *filament* density (not tubule density) is maintained in tapering

trees is that microtubule numbers are held constant in dendritic trees and therefore cannot serve alone as the shaping scaffold of tapering processes (although they can provide structural support).

Daughter-branch ratio This parameter is essentially one of diameter (daughter-branch ratio is a ratio of diameters); thus the arguments for the ultrastructural basis for segment diameter are applicable. The relevant question here is whether the cytoskeletal elements contribute to establishing the ratio itself. Although Weiss and Mayr (1971) determined the distribution of these components between daughter branches, they did not report the diameters of the branches, and their data do not shed any light on this problem. There is, in fact, reason to believe that ultrastructural elements contribute to the cytoskeleton as an underlying base for the daughter-branch ratio. This question bears further study.

Segment length and orientation The orientation and length of dendritic segments are both parameters for which there are, at this time, no direct ultrastructural correlates.

Developmental determinants of the fundamental parameters of shape

The ontogenetic process that gives rise to form in the CNS has been studied extensively for over 100 years (Boll, 1873). From more recent studies it has been proposed that the major factors operating during development fall into two classes, those controlled directly by the genome of each cell (intrinsic factors) and those arising from interactions between cells (extrinsic factors) (see Rakic, 1974, 1975; Berry and Bradley, 1976a,b; Lash and Burger, 1977). In neurons, *intrinsic factors* control cell division and the elongation of the cytoplasmic structures and membrane into elaborate ramifications. As shown in culture, isolated neurons and neuroblasts branch and form arborizations characteristic of neurons (Bray, 1970, 1973). *Extrinsic factors* arise from the specific interactive properties of glia (Guillery, Sobkowicz, and Scott, 1970; Rakic, 1971, 1974) and nerve cells. Of particular importance is the arrival of the afferent plexus (Morest, 1969a,b) and the subsequent formation of synapses (Skoff and Hamburger, 1974; Vaughn, Henrikson, and Grieshaber, 1974). In addition, fiber bundles, blood vessels, and brain surface areas passively influence the shape of developing neurons.

A major question here is whether correlations can be found between the parameters of form and the intrinsic or extrinsic developmental factors. Certainly shape parameters will not be determined entirely by intrinsic or by extrinsic factors but will, rather, result from their cooperative interaction. Nevertheless, certain parameters may be dominated by one or another of these factors.

Because consistent aspects of cell shape are likely to be controlled by the genome (for example, through the synthesis of subcellular components), one expects that cytoskeletal elements and those shape parameters that they influence will be determined largely by intrinsic factors. One approach to testing this reasoning is to search for invariances or constraints in shape parameters within cell types. On the other hand, variable parameters are more likely to be dominated by extrinsic factors and would reflect the lack of uniformity usually found in the environment immediately surrounding each developing cell. In this type of analysis one must keep in mind that a uniform external field (such as is found in the cerebellum) can be responsible for the consistent spatial orientation found in some cell types (for example, the planar character of the Purkinje-cell dendritic tree).

In pursuit of this line of thought I have included (when appropriate) the variance of the values obtained for the fundamental parameters of shape (see Table I). I would suggest that those aspects of cell shape that show the least variability (e.g., sum of cross-sectional areas of stem dendrites, terminal-segment diameter, branch power) are controlled by intrinsic factors and that those that vary widely (e.g., segment length and orientation) are controlled by extrinsic factors. Some support for this hypothesis can be found in the published studies of neuronal development that are described below.

Parameters Dominated by Intrinsic Factors

These and other investigations (Rall, 1959; Lux, Schubert, and Kreutzberg, 1970; Barrett and Crill, 1974a) have demonstrated that some parameters of form show little variability.

Although the size and shape of the soma are two of the most constant features of cells (Bok, 1936b, 1959) and represent the basis of cytoarchitectonic classifications (Campbell, 1905), the underlying ultrastructural basis for soma size is, to date, unknown.

Although neuronal processes are not as constant as the soma, there are three aspects that are relatively invariant and about whose ultrastructural basis some speculations can be made. The factors are (1) the sum of the cross-sectional areas of the stem processes, (2) the terminal-segment diameter, and (3) branch power. These three parameters of tree shape, and segment taper as well, are all measurements of di-

ameter and thus probably depend ultimately on the number of cytoskeletal elements such as neurotubules and neurofilaments whose synthesis and organization into a cytoskeleton is presumably controlled by the genome of each cell and little influenced by external factors.

Studies by Yamada and co-workers (1970) showed that microtubules were essential to neurite lengthening. When microtubule formation was stopped by application of colchicine, all processes failed to extend and some retracted. Thus the microtubules were believed to form a structural framework which was necessary for the elongation and stabilization of newly formed processes.

The role of microtubules in determining the three parameters listed above has not been addressed. One might speculate how the volume of a cell dendritic tree may be established genetically through limiting the number of tubules, filaments, or other cytoskeletal elements. In the simplest case, tubules alone are considered. One may propose that a cell generates a given number of tubules (composed of a fixed amount of tubulin) which continue to the terminals without dividing. If a constant tubule density is maintained, then regardless of the topology of the dendritic tree, the total volume of the tree is predetermined by the number of stem microtubules and their density. The conditions listed above are not unrealistic, for our results indicate that they are met in at least one cell type—the pyramidal-cell basal dendrites. In light of this suggestion, it is not surprising that the sum of the stem segments is correlated to soma size (Figure 3), since the stem-segment diameters merely reflect the number and density of their microtubules.

Parameters Dominated by Extrinsic Factors

There are four fundamental parameters of form that show considerable variability: (1) the number of dendritic stems emerging from the soma, (2) segment length, (3) segment orientation, and (4) daughter-branch ratio. These parameters are determined by the interactive influence (extrinsic) of properties of the interacting elements. Added to these influences are passive factors.

Stem diameter The diameter of individual stem segments and their distribution over the soma surface vary considerably from cell to cell. There is some support for the proposal that this feature is determined by the interaction of the developing cells with the environment. For example, although mammalian Purkinje cells show a marked potential to form more than one dendrite (as demonstrated by the appearance of numerous filopodia), adult cells have but one dendritic tree (Ramón y Cajal, 1911, 1929). This single tree may result from the strong influence of a climbing fiber (Kornguth and Scott, 1972). Following the capping of one pole of the soma by this afferent (Ramón y Cajal, 1911, 1929; Bradley and Berry, 1976), the only successful process extends from this pole. Subsequently all available microtubules and neurofilaments have but two pathways—the single dendrite and the axon.

In contrast, pyramidal cells migrate away from their strong afferent (and in doing this, away from the cortical surface) (Ramón y Cajal, 1929; Rakic, 1972). In this process the soma moves "downward," leaving the future apical dendritic process in its wake (Rakic, 1972). After arriving at its destined cortical level, multiple afferents interact with certain filopodia to direct the basal dendrites (see Schadé and Van Groenigen, 1961; Figures 4 and 5).

In the same way, cells developing in a completely isotropic field (e.g., the spiny cells of the caudate nucleus and central inferior olivary cells) have spherically radiating dendritic trees (Scheibel and Scheibel, 1955; Fox et al., 1971).

Segment length There is also some evidence from developmental studies that segment length is largely determined by extrinsic influences (Berry and Bradley, 1976b; Bradley and Berry, 1976). Developing cells express the potential to branch by producing numerous filopodia; however, the establishment of one of these mobile fingers into a stable process, with its own growth cone (see Tennyson, 1970), is dependent on its successful interaction with surrounding structures (Vaughn, Henrikson, and Grieshaber, 1974). Once process status is achieved, the length of each segment depends on the sites at which one or more filopodia are again stabilized as processes, and these then extend further, each with a growth cone. This continues until a mature neuron is formed. Thus the selection of filopodia determines the initial-segment length, the number of daughters at each branch point (Berry and Bradley, 1976b), and ultimately the orientation (see below) of each segment. These initial-segment lengths are not fixed (Berry and Bradley, 1976b), but can be altered as other developing processes enter the area (e.g., afferents and the dendrites of surrounding neurons). Under the influence of these processes the segment lengthens, largely, it seems, to provide room for these fibers. The interaction between the filopodia and the surround occurs through induction from synaptic afferents (Morest, 1969a,b; Vaughn, Henrikson, and Grieshaber, 1974) and by surface recognition of fil-

opodia by other surrounding structures (see Guillery, Sobkowicz, and Scott, 1970; Rakic, 1974, 1975; Berry and Bradley, 1976a,b).

Spatial orientation The spatial orientation of segments, and ultimately of the entire arborization, seems to be dominated by extrinsic factors. In this context the selection of filopodia determines the direction of successful processes (although certain preferences may be provided intrinsically). For example, the highly restricted directional growth of the Purkinje-cell dendritic tree (Figure 1) is largely controlled by its interaction with parallel fibers to form synapses (Rakic, 1974). Similarly, tissue-culture studies indicate that the achievement of the unique shapes of nerve cells, which we recognize as dendritic and axonal trees, are highly dependent on the external interactions of the cells during development. Although some studies show that there is a tendency for branching in the absence of specifically organized afferents (Pomerat et al., 1967; Privat and Drian, 1976) and that isolated cells form characteristic arborizations (Bray, 1970, 1973), forms typical of the area of origin of these cells do not occur unless their normal afferents are present (Wolf and Dubois-Dalaq, 1970; Privat and Drian, 1976).

Daughter-branch ratio Little work has been directed toward understanding the determinants of daughter-branch ratios. The selection of the dominant filopodium is determined by developmental interactions with the surrounding field (the influence of afferent fibers is especially marked: see Morest, 1969a,b; Berry and Bradley, 1976b), and this seems to be a promising direction in which to look for the determinants of this parameter.

Conclusion and summary

Utilizing computerized three-dimensional recording and analysis, I have described the shapes of arborizations in the nervous system by means of seven fundamental parameters. The basis of this model is that the stem and succeeding segments are successively divided by bifurcations until a limiting (terminal) diameter is reached. Thus stem diameter and terminal diameter are two fundamental parameters. A third parameter, segment length, is the distance between branch points. Three other parameters are related to branch points. First, branch power measures any change in the cross-sectional area across branch points. Second, daughter-branch ratio (the ratio of the diameters of the daughter branches) represents the distribution of cross-sectional areas between daughter branches. Third, the orientation of these segments is a measure of the angle and direction of each segment at the branch point. An additional parameter, segment taper, is needed in those neurons exhibiting a decrease in segment diameter between branch points.

Variations in neuronal form (characterized by differences in the shape of the soma and of dendritic arborizations) result from shifts in underlying intracellular structures. These structures compose a cytoskeleton that stabilizes elongation and provides a structural base for dendritic diameter. Microtubules are a fundamental component of this cytoskeletal core and extend, without branching, from their origin at the soma to the dendritic terminals. Furthermore, by maintaining tubule density, the total cross-sectional area (volume) remains constant from the stem through all the terminals. In some cell types this volume decreases toward the terminals. Here filaments in high density appear to increase the cross-sectional area at the base of the tree. This decreases the tubule density for the same region, yet the filament concentration remains constant (filaments are lost progressively as the total dendritic volume decreases distally).

During development, the sculpturing of this cytoskeletal core into arborizations takes place through an interplay between intrinsic influences provided through the subcellular structures and interactive influences that occur between the neuron and its surround. The following summary is presented to bring together some current thinking on neuronal form.

Three fundamental parameters (sum of the stem diameters, branch power, terminal diameter) show little variation and are believed to be primarily controlled by intrinsic factors during development. These parameters have a strong correlation with subcellular structures. Segment taper is probably also a member of this category. In contrast, individual stem diameter, segment length, daughter-branch ratio, and segment orientation all show significant variability. The developmental process that organizes the subcellular components into the cytoskeleton according to these four parameters is predominantly controlled by interactive (extrinsic) influences.

Soma volume reflects subcellular structures whose number is determined by the genome, which in turn constrains the total available volume of all dendritic structures of a cell. This volume is constrained by components of the cytoskeleton, basically tubules. The stem diameter of each dendrite is determined by tubule numbers, possibly from a total pool (in some cells filaments are an added factor). The distribution of this cross-sectional area between individual

dendrites on a soma is largely controlled by the "strength" of the interaction with surrounding structures. This selection process determines the number of tubules within each stem. The volume of the individual tree follows as a dependence on the numbers of tubules and filaments that compose the stem diameter while the length of segments distributes the volume.

In arborizations, the cross-sectional area of the stem is reduced by the generation of daughter branches. The first branch point defines the initial-segment length, which may increase with interspersion of additional neuropil (Berry and Bradley, 1976a). At each branch point the tubules of the core are divided between the daughters according to the "strength" of interactions (filopodial selection), thus giving rise to variations in the daughter-branch ratio. The branch power is determined by the properties of the core. For example, when tubules are the principal component, the cross-sectional area remains constant. If high concentrations of filaments are present, the cross-sectional area is reduced across the branch point and along segments. Finally, the branch point serves to determine the orientation for the subsequent daughter segments. This is also governed through filopodial selection (extrinsic factors). The bifurcation process is probably complete when the tubule number is reduced to a level such that a further division is insufficient to form two additional processes (each with at least a minimal complement). Further lengthening of terminal segments is possible even if division is not. (Thus terminal segments can be longer than other segments: see Berry and Bradley, 1976a,b.)

ACKNOWLEDGMENTS I am very appreciative of the expert technical assistance provided by S. Chen, S. Cuccio, and J. Gelbfish and the computer programming provided by M. Chujo and J. Gelbfish. Research was supported by USPHS grants HD-10934 from the National Institute of Child Health and Human Development and NS-13742 from the National Institute of Neurological and Communicative Disorders and Stroke.

REFERENCES

APÁTHY, S., 1897. Das leitende Element des nervensystems und seini topographischen Beziehungen zu den Zellen. *Mitth. Zool. Sta. Neapel.* 12:495–748.

BALTHASAR, K., 1962. Morphologie der spinalen Tibialis— und Peronaeus—Kerne bei der Katze: Topographie, Architektonik, Axon—und Dendritenverlauf der motoneurone und Zwischenneurone in den Segmenten L_6–S_2. *Arch. Psychiatr. Neurol.* 188:345–378.

BARRETT, J. N., and W. E. CRILL, 1974a. Specific membrane properties of cat motoneurons. *J. Physiol.* 239:301–324.

BARRETT, J. N., and W. E. CRILL, 1974b. Influence of dendritic location and membrane properties on the effectiveness of synapses on cat motoneurons. *J. Physiol.* 239:325–345.

BERRY, M., E. M. ANDERSON, T. HOLLINGWORTH, and R. M. FLINN, 1972. A computer technique for the estimation of the absolute three-dimensional array of basal dendritic fields using data from projected histological sections. *J. Microsc.* (Oxford) 95:257–267.

BERRY, M., and P. BRADLEY, 1976a. The application of network analysis to the study of branching patterns of large dendritic fields. *Brain Res.* 109:111–132.

BERRY, M., and P. BRADLEY, 1976b. The growth of the dendritic trees of Purkinje cells in the cerebellum of the rat. *Brain Res.* 112:1–35.

BERRY, M., T. HOLLINGWORTH, E. M. ANDERSON, and R. M. FLINN, 1975. Application of network analysis to the study of the branching patterns of dendritic fields. *Adv. Neurol.* 12:217–245.

BERRY, M., T. HOLLINGWORTH, R. M. FLINN, and E. M. ANDERSON, 1972. Dendritic field analysis: A reappraisal. *T.-I.T. J. Life Sci.* 2:129–140.

BIELSCHOWSKY, M., 1902. Die Silberimprägnation dei Axencylinder. *Neurol. Zentrabl.* 21:579–584.

BOK, S. T., 1936a. The branching of the dendrites in the cerebral cortex. *Verh. Kon. Med. Akad. Wetenshap.* 39:1209–1218.

BOK, S. T., 1936b. A quantitative analysis of the structure of the cerebral cortex. *Verh. Kon. Med. Akad. Wetenshap.* 35:1–55.

BOK, S. T., 1959. *Histonomy of the Cerebral Cortex.* Amsterdam: Elsevier.

BOLL, F., 1873. Die Histiologie und Histiogenese der nervösen Centralorgane. *Arch. Psychiatr.* (Berlin) 4:1–138.

BORN, G., 1883. Die Plattenmodelliermethode. *Arch. Mikrosk. Anat.* 22:584–599.

BRADLEY, P. M., and BERRY, M., 1976. The effects of reduced climbing and parallel fibre input on Purkinje cell dendritic growth. *Brain Res.* 109:133–151.

BRAY, D., 1970. Surface movements during the growth of single explanted neurons. *Proc. Natl. Acad. Sci. USA* 65:905–910.

BRAY, D., 1973. Branching patterns of individual sympathetic neurons in culture. *J. Cell. Biol.* 56:702–712.

BROWN, C., 1977. Neuron orientations: A computer application. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 177–188.

BROWN, P. B., 1976. *Computer Technology in Neurosciences*, Washington and London: Hemisphere.

CAMPBELL, A. W., 1905. *Histological Studies on the Localisation of Cerebral Function.* New York: Cambridge Univ. Press.

CHU, L. W., 1954. A cytoskeletal study of anterior horn cells isolated from human spinal cord. *J. Comp. Neurol.* 100:381–416.

CLENDINNEN, B. G., and J. T. FAYRS, 1961. The anatomical and physiological effects of prenatally administered somatropin on cerebral development in rats. *J. Endocrinol.* 22:183–193.

COLEMAN, P., C. GARVEY, J. YOUNG, and W. SIMON, 1977. Semi-automatic tracking of neuronal processes. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 91–110.

COLEMAN, P. D., and A. H. RIESEN, 1968. Environmental effects on cortical dendritic fields. I. Rearing in the dark. *J. Anat.* 102:363–374.

COLEMAN, P. D., M. J. WEST, and R. WYSS, 1973. Computer-aided quantitative neuroanatomy. In *Digital Computers in the Behavioral Lab.*, B. Weiss, ed. New York: Appleton-Century-Crofts, pp. 379–426.

COLON, E. J., and G. J. SMIT, 1970. Quantitative analysis of the cerebral cortex. II. A method for analyzing basal dendritic plexuses. *Brain Res.* 22:363–380.

COLON, E. J., and G. J. SMIT, 1971. A quantitative analysis of dendritic patterns in the cerebral cortex. *Acta Morphol. Neerl. Scand.* 9:21–39.

COLONNIER, M., 1964. The tangential organization of the visual cortex. *J. Anat.* 98:327–344.

DEITERS, O., 1865. *Untersuchungen über Gehirn und Rückenmark des menschen und der Saügetiere*. Braunschweig.

EAYRS, J. T., 1955. The cerebral cortex of normal and hypothyroid rats. *Acta Anat.* 25:160–183.

FOX, C. A., and J. W. BARNARD, 1957. A quantitative study of Purkinje cell dendritic branchlets and their relationship to afferent fibers. *J. Anat.* 91:299–313.

FOX, C. A., A. N. ANDRADE, D. E. HILLMAN, and R. C. SCHWYN, 1971. The spiny neurons in the primate striatum: A Golgi and electron microscopic study. *J. Hirnforschung.* 13:181–201.

FRIEDE, R. L., and T. SAMORAJSKI, 1970. Axon caliber related to neurofilaments and microtubules in sciatic nerve fibers of rats and mice. *Anat. Rec.* 167:379–388.

GARVEY, C. F., J. H. YOUNG, P. D. COLEMAN, and W. SIMON, 1973. Automated three-dimensional dendrite tracking system. *EEG Clin. Neurophysiol.* 35:199–204.

GELFAN, S., G. KARA, and D. S. RUCHKIN, 1970. The dendritic tree of spinal neurons. *J. Comp. Neurol.* 139:385–412.

GLASER, E. M., and H. VAN DER LOOS, 1965. A semi-automatic computer microscope for the analysis of neuronal morphology. *IEEE Trans. Biomed. Eng.* 12:22–31.

GLASSER, S., J. MILLER, N. G. YOUNG, and A. SELVERSTON, 1977. Computer reconstruction of invertebrate nerve cells. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 21–58.

GOLGI, C., 1874. Sulla fina anatomia del cervelletto umano. Reprinted in *Opera Omnia*, vol. 1: *Istologia normale, 1870–1883*. Milan: Ulrico Hoepli (1903), pp. 99–111.

GUILLERY, R. W., H. M. SOBKOWICZ, and G. L. SCOTT, 1970. Relationships between glial and neuronal elements in the development of long term cultures of the spinal cord of the fetal mouse. *J. Comp. Neurol.* 140:1–34.

HAGGAR, R. A., and M. L. BARR, 1950. Quantitative data on the size of synaptic end bulbs in the cat's spinal cord. *J. Comp. Neurol.* 93:17–36.

HILLMAN, D. E., M. CHUJO, and R. LLINÁS, 1974. Quantitative computer analysis of the morphology of cerebellar neurons. *Anat. Rec.* 178:375 (abstract).

HILLMAN, D. E., R. LLINÁS, and M. CHUJO, 1977. Automatic and semi-automatic analysis of nervous system structure. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 73–89.

HOOPEN, M. TEN, and H. A. REUVER, 1970. Probabilistic analysis of dendritic branching patterns of cortical neurons. *Kybernetik* 6:176–188.

KÖLLIKER, A. V., 1891. Die Lehrer von den Beziehungen der nervösen Elemente zu einander. *Anat. Anz. Ergazungshftr.* 1891:5–20.

KORNGUTH, S. E., and G. SCOTT, 1972. The role of climbing fibers in the formation of Purkinje cell dendrites. *J. Comp. Neurol.* 146:61–82.

LASH, J. W., and M. M. BURGER, eds., 1977. *Cell and Tissue Interactions*. New York: Raven Press.

LEVINTHAL, C., E. MACAGNO, and C. TOUNTAS, 1974. Computer-aided reconstruction from serial sections. *Fed. Proc.* 33:2326–2340.

LEVINTHAL, C., and R. WARE, 1972. Three-dimensional reconstruction from serial sections. *Nature* 236:207–210.

LINDSAY, R. D., ed., 1977a. *Computer Analysis of Neuronal Structures*. New York and London: Plenum.

LINDSAY, R. D., 1977b. The video computer microscope and A.R.G.O.S. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 1–19.

LINDSAY, R. D., 1977c. Tree analysis of neuronal processes. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 149–164.

LINDSAY, R. D., 1977d. Neuronal field analysis using Fourier series. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 165–175.

LINDSAY, R. D., and A. B. SCHEIBEL, 1974. Quantitative analysis of the dendritic branching pattern of small pyramidal cells from adult rat somesthetic and visual cortex. *Exp. Neurol.* 45:424–434.

LLINÁS, R., and D. E. HILLMAN, 1975. A multipurpose tridimensional reconstruction computer system for neuroanatomy. In *Golgi Centennial Symposium: Perspectives in Neurobiology*, M. Santini, ed. New York: Raven, pp. 519–528.

LLINÁS, R., and A. IBERALL, 1977. A global model of neuronal command-control systems. *BioSystems* 8:233–235.

LORENTE DE NÓ, R., 1934. Studies on the structure of the cerebral cortex. II. Continuation of the study of the ammonic system. *J. Psychol. Neurol.* (Leipzig) 46:113–177.

LUX, H. D., P. SCHUBERT, and G. W. KREUTZBERG, 1970. Direct matching of morphological and electrophysiological data in cat spinal motoneurons. In *Excitatory Synaptic Mechanisms*, P. Anderson and J. K. I. Jansen, eds. Oslo: Universitetsforlaget.

MCCONNELL, P., and M. BERRY, 1978. The effect of undernutrition on Purkinje cell dendritic growth in the rat. *J. Comp. Neurol.* 177:159–172.

MANNEN, H., 1966. Contribution to the quantitative study of the nervous tissue: A new method for measurement of the volume and surface area of neuron. *J. Comp. Neurol.* 126:75–90.

MOREST, D. K., 1969a. The differentiation of cerebral dendrites: A study of the post-migratory neuroblast in the medial nucleus of the trapezoid body. *Z. Anat. Entwicklungsgesch.* 128:271–289.

MOREST, D. K., 1969b. The growth of dendrites in the mammalian brain. *Z. Anat. Entwicklungsgesch.* 128:290–317.

MUNGAI, J. M., 1967. Dendritic patterns in the somatic sensory cortex of the cat. *J. Anat.* 101:403–418.

PALAY, S., 1956. Synapses in the central nervous system. *J. Biophys. Biochem. Cytol.* 2:193–201.

PALAY, S. L., and V. CHAN-PALAY, 1974. *Cerebellar Cortex: Cytology and Organization*. New York: Springer-Verlag.

PALDINO, A., and E. HARTH, 1977a. A measuring system for analyzing neuronal fiber structure. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 59–72.

PALDINO, A., and E. HARTH, 1977b. A computerized study of Golgi impregnated axons in rat visual cortex. In *Computer Analysis of Neuronal Structures*, R. D. Lindsay, ed. New York and London: Plenum, pp. 189–207.

PETERS, A., 1968. Characterization of microtubules, neurofilaments and cross bridges in various neuronal types. *Neurosci. Res. Program Bull.* 6:162–188.

PETERS, A., 1971. Stellate cells in the rat parietal cortex. *J. Comp. Neurol.* 141:345–374.

PETERS, H. G., and H. BADEMAN, 1963. The form and growth of stellate cells in the cortex of the guinea pig. *J. Anat.* 97:111–117.

PETERS, A., S. PALAY, and H. WEBSTER, 1976. *The Fine Structure of the Nervous System: The Neurons and Supporting Cells*. Philadelphia: W. B. Saunders.

PETERSON, T. S., 1955. *Analytic Geometry and Calculus*. New York: Harper and Brothers, p. 410.

POMERAT, C. M., W. J. HENDELMAN, C. W. RAIBORN, and J. F. MASSEY, 1967. Dynamic activities of nervous tissue *in vitro*. In *The Neuron*, H. Hydén, ed. New York: Elsevier, pp. 119–178.

PORTER, K. R., and L. G. TILNEY, 1965. Microtubules and intracellular motility. *Science* 150:382.

PRIVAT, A., and M. J. DRIAN, 1976. Postnatal maturation of rat Purkinje cells cultivated in the absence of two afferent systems: An ultrastructural study. *J. Comp. Neurol.* 166:201–244.

PURKINJE, J. E., 1837. Bericht über die Versammlung deutscher Naturforscher und Ärzte (Prag). *Anat. Physiologische Verhandlungen* 3:177–180.

RAKIC, P., 1971. Neuron-glia relationships during granule cell migration in developing cerebellar cortex. A Golgi and electronmicroscope study in *Macacus rhesus*. *J. Comp. Neurol.* 141:283–312.

RAKIC, P., 1972. Mode of cell migration to the superficial layers of fetal monkey neocortex. *J. Comp. Neurol.* 145:61–84.

RAKIC, P., 1974. Intrinsic and extrinsic factors influencing the shape of neurons and their assembly into neuronal circuits. In *Frontiers in Neurology and Neuroscience Research*, P. Seeman and G. M. Brown, eds. Toronto: Univ. Toronto Press, pp. 112–132.

RAKIC, P., 1975. Role of cell interaction in development of dendritic patterns. *Adv. Neurol.* 12:117–134.

RALL, W., 1959. Branching dendritic trees and motoneuron membrane resistivity. *Exp. Neurol.* 1:491–527.

RALL, W., 1962. Electrophysiology of a dendritic neurone model. *Biophys. J.* 2:145–167.

RALL, W., 1970. Cable properties of dendrites and effects of synaptic location. In *Excitatory Synaptic Mechanisms*, P. Andersen and J. K. S. Jansen, eds. Oslo: Universitetsforlaget.

RAMON Y CAJAL, S., 1909. *Histologie du système nerveux de l'homme et des vertébrés*, vol. I. Paris: Maloine.

RAMON Y CAJAL, S., 1911. *Histologie du système nerveux de l'homme et des vertébrés*, vol. II. Paris: Maloine.

RAMÓN Y CAJAL, S., 1929. *Studies on Vertebrate Neurogenesis*. Springfield, IL: Thomas.

REDDY, D. R., W. J. DAVIS, R. B. OHLANDER, and D. J. BIHARY, 1973. Computer analysis of neuronal structures. In *Intracellular Staining in Neurobiology*, S. B. Kater and C. Nicholson, eds. New York: Springer-Verlag, pp. 227–253.

SCHADÉ, J. P., and W. F. CAVENESS, 1968. Pathogenesis of X-irradiation effects in the monkey cerebral cortex. IV. Alteration in dendritic organization. *Brain Res.* 7:59–84.

SCHADÉ, J. P., and W. B. VAN GROENIGEN, 1961. Structural organization of the human cerebral cortex. I. Maturation of the middle frontal gyrus. *Acta Anat.* 47:74–111.

SCHEIBEL, M. E., and A. B. SCHEIBEL, 1955. The inferior olive—A Golgi study. *J. Comp. Neurol.* 102:77–132.

SCHMITT, F. O., 1968. The molecular biology of neuronal fibrous proteins. *Neurosci. Res. Program Bull.* 6:119–144.

SCHMITT, F. O., and B. B. GEREN, 1950. The fibrous structure of the nerve axon in relation to the localization of "neurotubules." *J. Exp. Med.* 91:499–507.

SCHULZE, M., 1871. Allgemeines über die Structurelements des Nerven Systems. In *Handbuch der Lehre von den Geweben*, S. Stricker, ed. Leipzig.

SELVERSTON, A. I., 1973. The use of intracellular dye injections in the study of small neural networks. In *Intracellular Staining in Neurobiology*, S. B. Kater and C. Nicholson, eds. New York: Springer-Verlag, pp. 255–280.

SHOLL, D. A., 1953. Dendritic organization in the neurons of the visual and motor cortices of the cat. *J. Anat.* 87:387–406.

SHOLL, D. A., 1955. The surface area of cortical neurons. *J. Anat.* 89:571–572.

SHOLL, D. A., 1956. *The Organization of the Cerebral Cortex*. London: Methuen.

SKOFF, R. P., and V. HAMBURGER, 1974. Fine structure of dendritic and axonic growth cones in embryonic chick spinal cord. *J. Comp. Neurol.* 153:107–148.

SMIT, G. J., and H. B. UYLINGS, 1975. The morphometry of the branching pattern in the dendrites of the visual cortex pyramidal cells. *Brain Res.* 87:41–53.

SMIT, G. J., H. B. UYLINGS, and L. VEEDMAAT-WANSINK, 1972. The branching pattern in dendrites of cortical neurons. *Acta Morphol. Neerl. Scand.* 9:253–274.

SMITH, D. E., 1973. The location of neurofilaments and microtubules during the postnatal development of Clarke's nucleus in the kitten. *Brain Res.* 55:41–53.

TENNYSON, V. M., 1970. The fine structure of the axon and growth cone of the dorsal root neuroblast of the rabbit embryo. *J. Cell. Biol.* 44:62–79.

TILNEY, L. G., and K. R. PORTER, 1967. Studies on the microtubules in heliozoa. II. The effect of low temperature on these structures in the formation and maintenance of the axopodia. *J. Cell. Biol.* 34:327–343.

UYLINGS, H. B., and G. J. SMIT, 1975. Three-dimensional branching structure of pyramidal cell dendrites. *Brain Res.* 87:55–66.

VAUGHN, J. E., C. K. HENRIKSON, and J. A. GRIESHABER, 1974. A quantitative study of synapses on motoneuron dendritic growth cones in developing mouse spinal cord. *J. Cell. Biol.* 60:664–672.

WALDEYER, W., 1891. Uber einige neurere Forschungen im Gebrete der Anatomie des Central Nerven Systems. *Berl. Klin. Wchnschr.* 28:691.

WANN, D. F., T. A. WOOLSEY, M. L. DIERKER, and W. M. COWAN, 1973. An on-line digital-computer system for the semi-automatic analysis of Golgi impregnated neurons. *IEEE Trans. Biomed. Eng.* 20:233–247.

WEISS, P. A., and R. MAYR, 1971. Neuronal organelles in neuroplasmic ("axonal") flow. II. Neurotubules. *Acta Neuropathol.* (Berlin) Suppl. 5:198–206.

WOLF, M. K., and M. DUBOIS-DALAQ, 1970. Anatomy of cultured mouse cerebellum. I. Golgi and electron microscopic demonstrations of granule cells, their afferent and efferent synapses. *J. Comp. Neurol.* 140:261–280.

WONG, W. C., 1967. The tangential organization of dendrites and axons in three auditory areas of the cat's cerebral cortex. *J. Anat.* 101:419–433.

WUERKER, R. B., and J. B. KIRKPATRICK, 1972. Neuronal microtubules, neurofilaments and microfilaments. *Int. Rev. Cytol.* 33:45–75.

WUERKER, R. B., and S. PALAY, 1969. Neurofilaments and microtubules in anterior horn cells of the rat. *Tissue & Cell* 1:387–402.

YAMADA, K. M., B. S. SPOONER, and N. K. WESSELLS, 1970. Axon growth: Roles of microfilaments and microtubules. *Proc. Natl. Acad. Sci. USA* 66:1206–1212.

ZENKER, W., and E. HOHBERG, 1973. A α-nerve-fibre: Number of neurotubules in the stem fibre and in the terminal branches. *J. Neurocytol.* 2:143–148.

ZENKER, W., R. MAYR, and H. GRUBER, 1973. Axoplasmic organelles: Quantitative differences between ventral and dorsal root fibres of the rat. *Experientia* 29:77–78.

4. Applications III--Medical Imaging

A menu-driven user interface for a physician's imaging console

Jean-Bernard Massicotte, Ronald E. Wurtz, Richard W. Benster, E. Klingenberg

Contour Medical Systems, Inc.
1931-A Old Middlefield Way, Mountain View, California 94043

Abstract

Manipulation of several sets of two-dimensional cross sectional slices of computerized tomography (CT) or magnetic resonance (MR) data and the mental integration of such a large volume of information are major problems encountered by radiologists and surgeons in their attempt to make diagnoses. Contour Medical Systems has developed a physician's imaging console and a user-friendly interface intended to make this information immediately available and to facilitate image analysis. In this paper we will focus on the development of this user interface.

Introduction

With the development of increasingly powerful medical imaging scanners, an urgent need exists for providing radiologists with more support to process the large volume of generated data (either CT or MR). Contour Medical Systems has developed a workstation, the CEMAX-1000, which alleviates this lack of support and responds to physicians' needs to store, archive, view, process, and analyze image data. The major innovation of this tool is the integration, into one system, of a high-resolution computer graphics system, a menu-oriented interface using a digitizer tablet for cursor control instead of a keyboard, and a large bank of high-level functions intended for extensive image analysis. Substantial effort was put into the development of this easy-to-use interface, wherein each function is initiated exclusively by cursor selection from color menus.

Basic concepts of user interfaces

The standard interface in medical imaging consoles uses special keypads for the selection of menu items, a keyboard for text input, a trackball or joystick for control of cursors on images, and one or two medium-resolution grey scale screens for display. Outside the medical world, image processing systems using high-resolution (up to 1024 lines x 1280 pixels) color graphics displays have been well established for a long time. However, functions are still selected via keyboard commands of varying complexity with the associated problem of memorizing the commands. Moreover, dedicated function keys, as used in both medical and CAD/CAM fields, either impose restrictions on the number of functions available per menu or lead to complex multi-key sequences for one selection(1).

Completely flexible interfaces, with all control done via menu-buttons on the screen, became popular with the introduction of systems such as Xerox STAR (2) and Apple LISA. In these systems, the logical sequence of menus, called a menu tree, is not bound by any physical keypads and can be of arbitrary complexity. Menu items are selected by moving a cursor on a desired menu-button, using an input device such as a mouse, a digitizer tablet and stylus, a touch-sensitive screen, or a similar locator device. In the case of the mouse, as is used in the Xerox STAR and Apple LISA, pressing a button on the mouse itself signals the selection of the menu item.

These flexible menu trees paved the way for easy-to-use interfaces. The organization of the menu tree, and of each menu, as well as the text or icon related to each menu-button, are designed to provide explicit structural information of the system to the user. With such a self-explanatory interface, the user requires little or no training, which is considered a prime advantage of menu-driven systems(3).

These emerging concepts of easy-to-use menu driven systems, together with the special problems in the manipulation of large amounts of image data, underlie many of the design decisions in the CEMAX-1000 user interface. Some of the overall decisions will be outlined here, with details presented in subsequent sections.

The most important decision in the design of the user interface was to implement only one interactive input device, the tablet, and one interactive output device, the screen (see Figure 1). The tablet is used jointly with a puck instead of the standard stylus. Although similar to the mouse in appearance and functionality, the puck has a small coil which generates a magnetic field detected by the tablet, giving absolute position with respect to the tablet, while the mouse, via its rolling balls, gives a position relative to its initial arbitrary position.

Tablet and mouse are equally efficient as an input device(4). However, at the time the system was designed, the digitizer tablets available seemed more reliable than the mice, eventhough this point is now debatable. It was the only motivation behind the selection of the tablet; either could be used.

The combination tablet/puck because of its accuracy, its simplicity and its full flexibility in terms of freedom of motion, makes the selection of menu items fast and easy, which is an important factor in the design of a good user interface(4). These properties of the tablet made it preferable to the trackball and the joystick. Finally, touch-screens were rejected for several reasons: arm-fatigue caused by reaching for the screen during long viewing sessions, the low spatial resolution of the finger when used as a pointing device on the screen, and the difficulty of drawing on a vertical surface like a screen.

A draw-back to the current design, a consequence of the interface's simplicity, is the lack of a totally flexible text input facility for keeping personal comments or annotations to image files. Currently, basic annotations to certain images are possible by using menus with selectable keywords (for example, images reconstructed in arbitrary planes are classified with selectable words such as "saggital" and "coronal", chosen from a menu). This text input function lacks flexibility because of its limited vocabulary and its limited scope. However, its simplicity is appealing and its efficiency is easily improved by use of a larger bank of words.

Several techniques are available to provide a more flexible text input facility. One makes use of graphic representation of a keyboard for a selection of alphabetic characters via the tablet. A keyboard can also be implemented by means of a voice recognition device, where the letters of the words are spoken. Word recognition is faster but again, the vocabulary is limited. The last solution is the traditional keyboard. However, its use should be limited to text input exclusively, with any kind of program control being reserved to the menus, because the keyboard is an unfriendly input device to the untrained user.

The second important decision in the design of the user interface was the selection of the single screen, instead of the dual screens common in medical imaging consoles, in order to minimize eye-movements and to simplify hand-eye coordination. The screen has a high-resolution rectangular format (1024 lines x 1280 pixels) and is partitioned into two regions. A square area (1024 lines x 1024 pixels) is used for image and patient information. A narrow strip of this image area is reserved for display of a metered bar, for interactively setting various parameters such as image contrast, viewing angle for 3-dimensional display, etc. The second region is a strip (1024 lines x 256 pixels) on the right, used for menus and basic instructions, where cursor hits are used for menu selection. While moving to deeper layers of the menu tree, the user goes through a sequence of nodes. The menus are laid out to show the current menu (ie a list of selectable commands) as well as this sequence of higher-level nodes, giving the user the freedom to backtrack easily to higher layers at any point. An on-line help facility is also available on each menu through a help button, giving basic instructions for the functions currently available. Figure 2 shows an example of the menu, with selectable functions, help button,and two menu-buttons for access to higher levels.

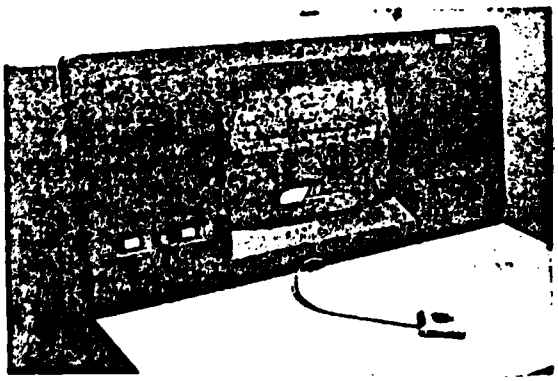


Figure 1. The CEMAX-1000 workstation The puck is used jointly with the tablet to control the screen cursor.

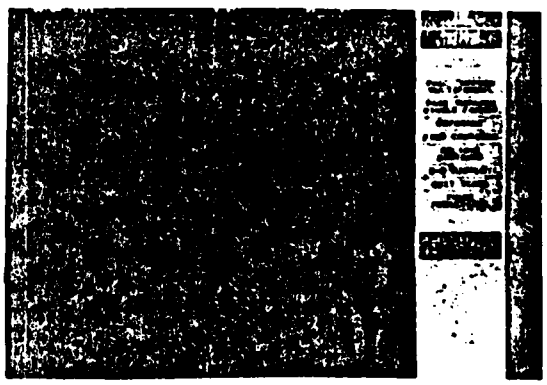


Figure 2. Example of a menu with 2 pop-up buttons at the top, the title of the menu, the currently available functions and the help button.

Lastly, the design required the selection of formats for data presentation in the image and text area, to be used across all the major functions. For example, when only textual information is to be presented, the entire 1024 x 1024 window is a scrollable text display. When viewing text lists, cursor hits are used to select an item from these lists, for example, to select a patient's data set. For viewing images, the image area can be used as as collage of small or large windows, depending on the user's need for a quick overview of several images or a more detailed study, one image at a time. When viewing images, the cursor hits in the image are used for initiating numerous image analysis functions.

Access to the data base

CT image data stored with standard format on magnetic tapes are the data input to the workstation. Currently, magnetic tapes coming from most major scanner manufacturers can be read by CEMAX-1000. The reading of the tapes as well as any subsequent storage operations like deletion of data sets of images, archive and restore from/to cassette, is inititated from a menu button. All internal data management is automatic and invisible to the user.

One the data are read, the user can display a scrollable text list showing all patients currently on the disk as well as some information relevant to each patient such as run number, the time the scan was done, and number of images stored per patient. The user selects particular data sets for data analysis by highlighting numbers of this list through cursor hits. The selected members are then dsiplayed in reverse video (see Figure 3). Unlike most other systems, there is no need for a keyboard to enter a file access number, making the selection faster and virtually error free.

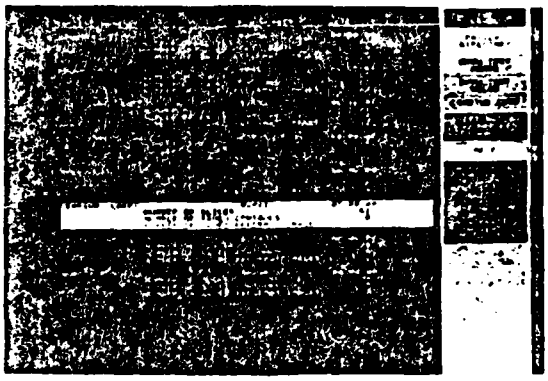


Figure 3. Patient directory showing a selected data set. The selection is done via cursor hit.

Use of the projection-views for slice selection

On most systems, including the CEMAX-1000, selection of a data set is followed by a quick viewing of some slices in order to locate a tumor, an organ, a fracture, a malformation, or whatever motivated the scanning of the patient. The radiologist may have to step through many slices, without knowing exactly where the slices are with respect to the whole volume of data. To help this search, conventional consoles sometimes display a so-called "reference image" (like a SCOUT, used by GE, or the SHADOW-VIEW used by Technicare for their MR images); annotations on this reference image show the location of the slices. However, since the reference images are generated prior to the scanning operation, they do not correspond to the true data set and are not the best representation of the volume of data.

In contrast, every time a patient's data set is read or edited, the CEMAX-1000 automatically creates, using the original data, its own reference images, called PROJECTION-VIEWS. The projection-views are digital reprojections of the original slices meant to simulate conventional antero-posterior and lateral x-ray images (see Figures 4 & 5). They are done by projecting the intensity values of the whole data set onto a plane, similar to real x-rays travelling from the source, through the body, to a detector plane. A proper scaling with some interpolation is done longitudinally in order to give true proportions (since the set of slice data is usually not isotropic), followed by some contrast enhancement to improve image quality.

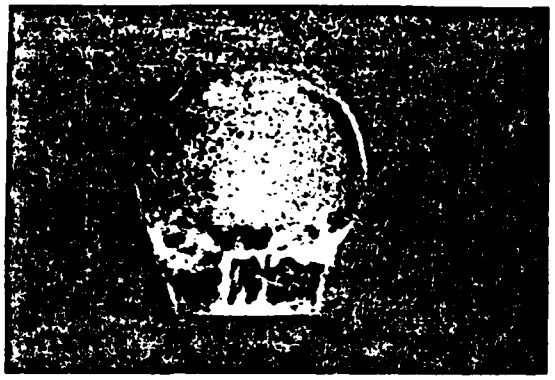


Figure 4. Projection-view showing an antero-posterior view of a skull.

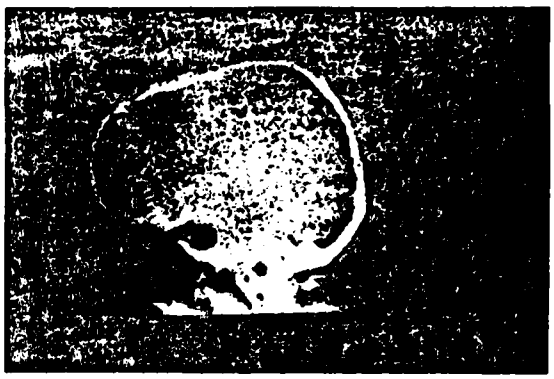


Figure 5. Projection-view showing a lateral view of a skull.

The projection-views are useful for a quick visualization of the entire data set, showing the spatial extent of the data, a silhouette of the patient with some internal structures (such as bony structures and sometimes organs and tumors), and also bad slices if any (like transversally shifted slices due to patient motion). They are also useful for zeroing-in on the slices of interest: once the physician has located a particular structure or organ on a projection-view, he can select a plane via cursor hit, resulting in the automatic display of the corresponding slice. The functions "up one slice" or "down one slice" are then used for fine tuning. Up to 8 slices can be displayed at the same time on the screen, using this kind of interaction (see Figure 6).

Furthermore, by rotating the cursor line on either the projection-views or on a transverse slice, the user can define a plane for creation of a reformatted image. Through this interaction, possible orientation of the cutting plane are parallel to the longitudinal axis, (or cephalo-caudal axis), parallel to the lateral axis, or parallel to the antero - posterior axis. A sagittal cut is obtained with a longitudinal line on the antero-posterior view; a coronal cut is obtained with a longitudinal line on the lateral view (see Figure 7); an "oblique" cut is obtained with a tilted cursor line on a projection view or on a transverse slice (see Figure 8).

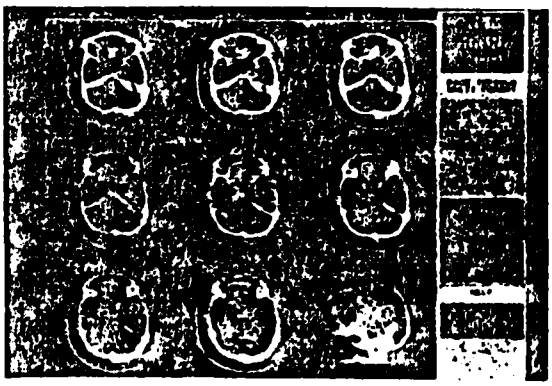


Figure 6. The "MULTI-FORMAT" menu, with 8 slices displayed simultaneously. Moving the cursor line on the projection-view results in the automatic display of the corresponding slice. The 2 lines displayed on the projection-view show the range of slices displayed.

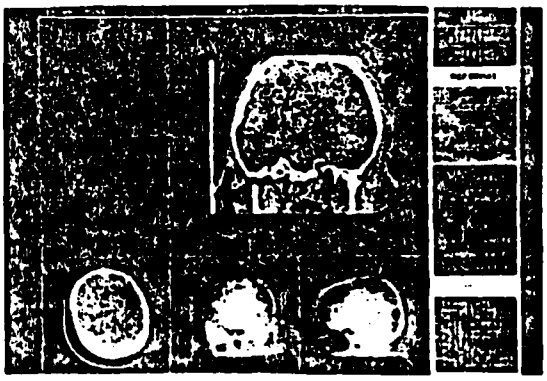


Figure 7. The "REFORMAT" menu, with a coronal cut of the skull. The cursor line on the projection-view is used to define the plane of the reformatted image.

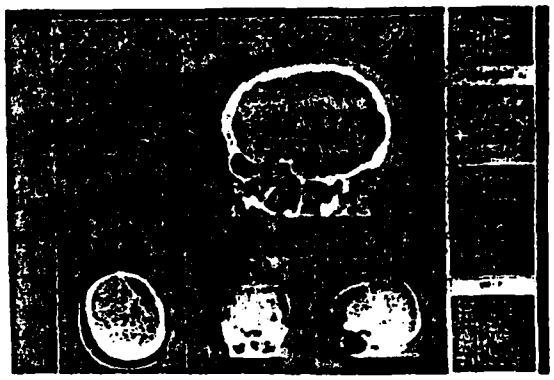


Figure 8. The "REFORMAT" menu, with a cut done at an oblique angle. The cursor line on the slice, in the lower left corner, define the plane of the cut.

Other interactive tools for image analysis

Once a transverse slice or a reformatted image is displayed, the user can move on to more extensive image analysis by using some of the interactive tools available in the system, such as tissue highlighting.

For tissue highlighting, the user selects interactively the width of the window of intensity values to be displayed on the screen, as well as the level of this window. This interaction is done by moving the cursor along the metered color bar. Grey scale as well as color can be used, depending on the user's preference.

By defining a region-of-interest on an image, and by picking a destination window, both via cursor hits, the user can magnify any portion of the image if he needs to look at small details. Similarly, he can select pieces of images and display any subset of them simultaneously, to facilitate the viewing of pertinent images. Another feature allows the user to magnify any portion of an image to life size, sometimes essential for planning surgery.

The CEMAX-1000 also provides a set of meaurement tools. For example, after having located a lesion or other relevant structure, the physician might be interested in making some size measurement, either for pre-operative planning or simply for diagnosis. He can overlay a grid on any transverse slice or any reformatted image for a quick evaluation of sizes. The exact distance between any 2 points is obtained simply by selecting two points via cursor hits.

Volume of structures are also available if "surface extraction" of desired tissues (lesion tissues or bony tissues for example) has been previously performed on those structures. The first purpose of this surface extraction operation is to reconstruct a three-dimensional solid model of an object by finding its outer surface, or contour. Figure 9 shows examples of 3D images obtained using this technique.

The contour-finding algorithm is automatic, but requires the user to define some initial parameters interactively. First, with the "tissue highlight" function, the user must define the range of intensity values (in Hounsfield units) corresponding to the object to contour. The algorithm uses intensity thresholding to separate the desired tissues from other tissues. Second, the user can define the volume of data to process, instead of using the default data set. Restricting the volume to a box fitting only the desired object is essential to avoid contouring undesirable tissues. The "slice-range" function and "region-of-interest" function are used for this purpose. A contour coming from a slice is stored as a series of linked vectors. The resulting set of contours is saved in a contour file for subsequent processing and three-dimensional display.

Unfortunately, extraction of a structure is not always succesful when neighboring tissues have similar intensity values. To alleviate this problem, the user can define an irregular region-of-interest (as opposed to the rectangular region-of-interest), by drawing on each slice, via the cursor, a region delimiting the area to process. Thus, flexibility is gained at the expense of not being able to do automatic extraction over the whole slice range.

Because of the accuracy of the 3D solid models generated by extracting surfaces from the transverse slices, sets of contours have subsequently been used to machine the models from synthetic material for pre-operative planning and for implant purposes. A molding process with a two-part mold is used to create life size models of the original object. The two half-molds are machined from a wax substrate then assembled and filled with a resin-based polymeric material, or a bio-compatible substance (if intended for an implant). The mold is removed once the inner material is solidified, showing an accurate reproduction of the object (see Figure 10).

Another feature of the CEMAX-1000 allows the user to contour a series of uniformally separated reformatted images, instead of the original slices (see Figure 11). Contouring the data at an angle is essentially equivalent to a rotation of the object in space. sometimes useful to observe the object from angles impossible to reach using the standard contours. The other and quite important benefit of using reformatted images for contouring is the creation of an isotropic data set, resulting in smoother surfaces. The higher quality of the 3D images fully justifies the longer processing time introduced by the reformatting step and by the resulting larger number of slices to contour.

Figure 9. 3D images of hips and femur of patient with congenital hip dysplasia. The metered bar on the left is used to define the orientation of the object.

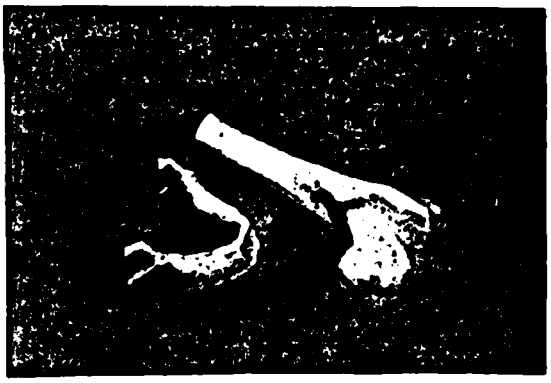


Figure 10. Milled models of portion of a mandible and a femur. The mandible is used to design a subperiostal implant. The femur is used for pre-operative planning.

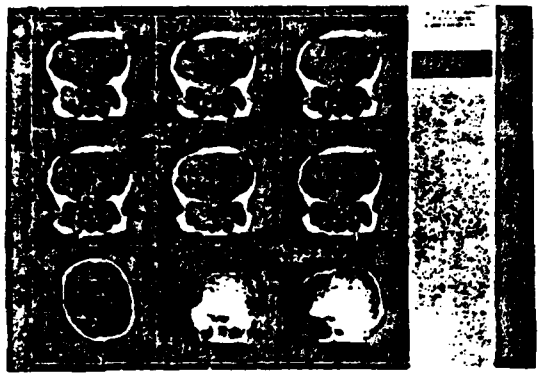


Figure 11. Contouring of a series of coronal images. The box on the lower left image defines the volume of data to be processed while the horizontal line shows the position of the current coronal image. The bright regions on the reformatted images correspond to the extracted tissues, in this case, bone.

Conclusion

CEMAX-1000 was designed to provide physicians with a friendly and easy-to-use workstation for manipulation of large volumes of data and for extensive image analysis, manipulation and analysis not possible or easily done on existing medical imaging consoles. The key elements of CEMAX-1000 user interfaces are the combination digitizer tablet and screen cursor used for all command input, the high-resolution screen for all display output, the menu-oriented control, and the logical and easy-to-follow sequence of events.

Other distinguishing characteristics of the CEMAX-1000 system include the selection of a patient's data set by highlighting members of a scrollable text list, thus eliminating the need for a standard keyboard. Also included is the ability to create with the original slices a set of sophisticated reference images, the "projection-views", to access slices in the data-base. These projection-views give quick visualization of the true data set, facililitating access to a particular slice or a reconstructed image in an arbitrary plane, as well as facilitating the definition of a volume of data for the contour-finding algorithm. Further interactive analysis can be performed using tissue highlight with grey scale or color, copy and magnify of slices or reformatted images, distance and volume measurement, and reconstruction and display of solid objects.

References

1. PDC Operator's Guide, General Electric CT Scanner, 1982.
2. Canfield Smith, D., Irby, C., Kimball, R., Verplank, B., Harslem, E., "Designing the Star User Interface", Byte, Vol. 7, No. 4, Apr. 1982, pp. 242-282
3. Grimes, J.D., "A cognitive view of user interfaces-draft," SIGGRAPH '84 tutorial, "Psychology of User-Computer Interfaces, "No. 4, Jul. 1984.
4. Doley, J.D. Wallace V.L., Chan P., "The Human Factors of Computer Graphics Interaction Techniques," IEEE Computer Graphics and Applications, Vol. 4, No. 11, Nov. 1984, pp. 13 - 45.

AUTOMATED MACHINING OF CUSTOM ANATOMICAL MODELS USING A SMALL SCALE INTEGRATED FACILITY

John C. Vogel, Ph.D.
CONTOUR MEDICAL SYSTEMS

Abstract

The design and manufacture of models of anatomical objects in a small-scale, low cost, automated machining center is accomplished using data acquired via computed tomography scanning. The data is thresholded and contoured to produce spatially organized polygons representing three dimensional shapes. The polygon boundary representations are the database for interactive modification using computer aided design (CAD) tools, two and three dimensional display and automatic conversion to tool path commands. The set up and implementation of a small computer integrated manufacturing (CIM) center to profitably produce complex custom 3-D parts requires careful consideration of data representation, communication, and verification issues to effectively handle the high data throughput.

AUTOMATED MACHINING OF CUSTOM ANATOMICAL MODELS USING A SMALL SCALE INTEGRATED FACILITY

Introduction

Recently, orthopedic and plastic surgeons have begun to use models of patient anatomy and custom designed implants to plan surgical procedures and perform reconstructive surgery (Ref. 1 - 4). The use of models decreases the uncertainty, time, cost and patient trauma required in those surgical procedures. For example, models are used in pre-operative planning to determine the placement of specific cuts or to select among different prosthetic devices. In another instance, the operation to take a dental impression, prerequisite to making a mold and from that an implant, can be eliminated. Any of the various computed tomographic scanning devices can supply the three dimensional data used to make corporeal models. Digital scanning (either x-ray or magnetic resonance) is becoming more prevalent; there are several thousand installed scanners in the United States alone.

In a typical application, the region of interest in a patient is scanned using one of two digital techniques, that of x-ray computed tomography (CT) or nuclear magnetic resonance (NMR). The output of either of these methods is a relatively large data file, on the order of 10 to 20 megabytes, containing, typically, 50 cross-sectional slices of data (Fig. 1) with one to two millimeter spacing and pixel size. Each pixel value represents the density of the tissue integrated over a small volume (voxel) of space.

Typical slice of computed tomography data

Fig. 1

Post-processing of this data is important to extract all the information useful to the physician. One company producing such post-processing equipment is Contour Medical Systems, Inc. (CMS) which markets an imaging console, the Cemax 1000 (Fig. 2). This imaging station reads in and selectively thresholds the data on a slice-by-slice basis. Edge detection algorithms generate contours files from the thresholded data. The individual contours are combined and used as the basic data structure for the display of various views of the reconstructed three dimensional object (Fig. 3). In addition to obtaining the object description from automated scanning, the physician can design arbitrarily shaped objects using the supplied CAD software, for example, the models and implants shown in Fig. 4.

CEMAX-1000 imaging console
Fig. 2

Three-dimensional images of reconstructed skull
Fig. 3

Samples of various models and implants made using the CEMAX-1000
Fig. 4

There is a variety of different options for displaying the scan data and any edited or created objects. First, the individual scan slices can be viewed showing a particular cross-section of the region of interest. The contours found by thresholding and doing the edge detection can also be viewed on a slice-by-slice basis. Equally as important, the contours can be combined together to form a three dimensional representation of the part and displayed in 3-D (ringstack). Alternatively, an image of the object can be displayed using depth encoding where the brightness of a pixel is proportional to the distance from the observor. Shaded images using real-time light source can also be displayed.

The availability of this boundary representation database naturally suggests the manufacture of physical models of the desired region. Due to the arbitrary shapes involved, the total custom nature of the models and the relatively large amount of data, it was necessary to maximize the degree of automation and "automatic processing" of the data in the design and manufacturing process. This paper presents a description of the manufacturing process as it stands today, followed by a discussion of the technical issues and engineering trade-offs necessary to achieve an economic benefit from the manufacture of "quantity one" complex three dimensional parts. Management questions, such as recruiting technical people versus hiring consultants, are not discussed. Particular emphasis is placed on the systems issues for a low cost (<$300,000), small scale CIM facility.

4.2.5

System Description

The computer integrated facility developed to support the objectives of low cost and rapid turnaround is discussed in this section. This includes details of the specifics of the design process, manufacturing equipment and software; discussion of why this particular configuration was chosen is reserved for the following sections.

All data processing is done on the Cemax-1000 system, a Multibus based computer with a Motorola 68000 central processing unit and Unix operating system. The system includes a 9-track tape drive, 160 megabyte streaming tape cartridge for archiving, three RS-232 ports and the imaging console. The console is a high resolution color graphics device (1024 x 1280 x 10) with a puck for user input via menus. Up to three standard RS-232 terminals can be attached to the ports, or as in the process presented here, a numerically controlled milling machine can be connected in place of one of the terminals. Because the Unix operating system is multi-tasking, the Cemax-1000 can, for example, simultaneously drive the imaging console, a terminal and the milling machine. The mill is a Bridgeport R2E3 with three translational degrees of freedom.

The entire process is menu-driven via the puck and on-screen menus. The data is read from the tape, thresholded and contoured, and displayed using the buttons on the puck to make menu selections. Several alternative algorithms are available for viewing the resulting contour files. The individual contours can be shown (or modified), the contours can be stacked and viewed ("ring stack"), and shaded or depth-encoded images can be produced. Modification of the contours is performed using the puck, the edited contour file can be redisplayed at any time for visual verification of the changes. The polygons are edited using typical CAD functions; an important option for implant design is the "mirroring" of individual polygons or groups of polygons about a user defined plane. This function is used when there is damage to only one side of the body. The mirroring capability helps the designer achieve a better fitting and more aesthetically pleasing part by using the undamaged side of the body as a template for the damaged side.

Further verification of the designed prosthetic is done using the metrics embedded in the software which provide, for example, measurements of angles and distances between points. Once the contours have been generated and verified, both quantitatively and visually, the process moves away from continuous user interaction and becomes self-automated. The present system produces models using two-part molds, and so requires two surfaces, the matching halves of the mold. The Cemax-1000 outputs a pair of visible surfaces to disk to form the two halves of the mold. This new way of automatically producing a tool path is discussed in more detail in a later section. At this point, the contours are converted from a closed polygon representation to a 2-D array format where the values of the entries in the array represent the depth of the surface.

Post-processing of the tool path is carried out in two steps. The first is to smooth the data, eliminating the effects of the discretization due to digital scanning techniques; the second is to generate a mill file from the smoothed data suitable for output to the Bridgeport mill. The smoothing is done using a weighted-average filter which operates over a small neighborhood of three points. The visible surface file is viewable on the Cemax-1000, with limited user interactivity, for verification of the results of the smoothing operation. The smoothed data arrays are then processed to produce the mill files themselves. Milling can require more than one pass, and, optionally, different sizes and shapes of tools. Interference checking is performed at this point using, in part, an approximation to the surface normal. The tool path itself can be displayed graphically to verify the result of the intervening step (Fig. 5).

A tool path generated automatically using the visible surface algorithm

Fig. 5

A typical mold in the process of being milled

Fig. 6

The Bridgeport mill is driven via a RS-232 link at a serial rate of 9600 bits/second to mill the two mold halves out of machinable wax. (Fig. 6). The size of the data files, typically 1/2 to 1 megabyte, and the large number of straight line segments typically 20,000 to 40,000, necessitate rapid handling of the milling commands. The data is sent out in the form of a series of Cartesian coordinates triplets which the Bridgeport interpolates with linear segments. The milling out of the two mold halves taken anywhere from 5 to 50 minutes depending on the size and complexity of the part and the spacing of the original scan data. Most objects require between 20 and 30 minutes, per half mold. Indexing holes are manually added by the operator. Following the milling of the two halves, the mold is assembled and a two-part self-curing polyethylene model material is mixed and poured (Fig. 7). Alternatively, the anatomical part itself (the positive) can be milled out of any of the available biocompatible materials.

A mold and part after milling and forming

Fig. 7

Issues in Implementing "Paperless" Design and Machining Centers

The following sections address some of the fundamental issues revelant to the manufacture of complex shapes using extensive computer integrated manufacturing. The biomedical application outlined above is used as an example. The specific numbers cited are based on our experience and are revelant only for this particular application. Actual throughput of data and parts production rates are given in the last part of this discussion.

Four topics are discussed: computer hardware requirements, database integrity and verification, data representation and management, and the actual model throughput achieved in practice. The second and third topics are primarily concerned with software development and use.

I. Hardware Requirements

Central processing unit requirements can often be met more efficiently by a combination of a relatively low-cost "central" processor and the appropriate co-processors, than by a single large expensive general purpose processor. The CPU in this case acts more as a central switchboard than as a classical computing engine. For instance, super micros based on 32-bit microprocessors can perform many of the housekeeping chores while number crunching is done on relatively inexpensive numeric processors with price/performance ratios surpassing 1 million floating point operations per second (1 MFLOPS)/$1000. Disk input/output (I/O) can be offloaded to "smart" disc controllers, decreasing effective disk access time by caching or other techniques. Increasingly sophisticated graphics processors which handle many of the "higher" level functions, such as shading, reduce the need for a powerful, but expensive, central processor.

Distributed systems tend to be bus-based. The primary advantage is flexibility; the main disadvantage is the limitation due to the bus bandwidth, the speed with which data can be moved through the bus. However, new bus architectures with higher data throughputs, equivalent to that of present day minicomputers, are becoming available. For instance, a 68000-based system running on a 10 MHz clock can support enough bus throughput, memory and peripheal devices to simultaneously run several terminals, one milling machine, and a single graphics device. For numeric processing, a 15 MFlops, $15,000 array processor is available. In the future, more multiprocessor and true parallel processing systems will become available, one of the objectives of current systems must be to retain the flexibility to incorporate such advances; bus-based systems offer one such way.

System resources necessary for a complete, but low cost machining center include, beside the CPU and mill itself, core memory, the hard disk(s), graphics subsystem, archiving and communication tools. Core memory requirements are somewhat offset by a fast hard disk and a virtual memory operating system, which treats the hard disk as part of core. However, performance deterioration is noticeable when core memory drops below the amount needed to maintain the program and associated data files in memory, (2 - 3 megabytes in the Cemax-1000) due to the greatly increased disk I/O required. The hard disk system must have enough throughput and capacity for 4 to 5 in-process designs (at least 100 megabytes, excluding operating system and software overhead).

The hard disk and system bus must be able to display images relatively quickly. For instance, a hard disk transfer rate of 1 megabyte/ second results in a one second display time for a typical 1024 x 1024 x 8 high resolution display assuming, unrealistically, no other current processes. The CPU and graphics subsystem must be able to maintain sufficient speed to the graphics display and puck to update the console at an ergonomic rate. This does not usually present problems with 2-D based systems, but becomes more difficult and expensive with 3-D. A rough ~~out~~guide line is that user I/O activity must occur every few seconds.

The archiving of data is an under appreciated task until a user can not find some "old" data. The primary issues are cost, access time and reliability. Magnetic tape, in the form of a 1/4" streaming cassette tape, each of which store 67 megabytes, is one fairly low cost solution. This medium has the advantages over 9-track tape of speed, more compact storage, and lower cost per megabyte. This solution is appropriate where data sets are on the order of 10 - 20 megabytes, not too many data sets are on each tape and any data set can fit on just one tape. Retrieval time is on the order of 10 - 20 minutes. For much larger data sets, the storage medium should be scaled to fit a small integer number of files per tape. For very large (greater than 100 megabytes) data files, optical disk technology supplies write-once, read optical disks with one Gigabyte storage capability. This solution is presently more expensive than the mature tape storage technology.

The manufacture of complex 3-D objects requires large data files necessitating transfer by all electronics means to increase reliability, throughput and reduce clerical error. This requirement also includes the milling machine instruction files which are too large for facile handling by paper tape since each part is to be machined only once. In practice, all electronic communication means using either a central facility controlling all the hard disks, archiving devices, and milling stations, or a decentralized facility using multiple independent smaller systems connected via a high speed network, such as Ethernet.

For the custom medical prosthetic application outlined in the system description, the second alternative was chosen for several reasons. First, the initial capitalization costs were much lower starting with a reasonably inexpensive Multibus system. Second, the in-house capability can easily be expanded by adding more systems as necessary and interconnecting them with the Ethernet network. Expandability can include any graphics device, specialized processor or milling machine that can communicate via Multibus, Ethernet or RS-232C protocols, a very large fraction of all such equipment. Further, it was recognized that it is inefficient use of CPU cycles to share a single processor among slow terminal I/O and high demand disk access and graphics jobs. Small, cheap CPUs should perform terminal and mill I/O, specialized processors should perform disk I/O, graphics and numerical processing.

II. Database Integrity and Verification

Maintaining database integrity and verification is defined as keeping errors out of the database, either factual or conceptual, by periodic checking of the contents of the database. This is obviously a very important concept, particularly where many different people and/or programs are changing the database. As more and more of the manufacturing process is sped up and automated the potential increases for fast propagation of errors, with a proportionally greater chance of misuse of resources. In addition to the checks built into any system to ensure the prevention of the introduction and propagation of errors, specific concepts must be incorporated into a CIM system to deal with the computer issues. Designing an integrated manufacturing system to prevent, detect amd correct errors includes: removing human intervention from the data stream. designing software so that illegal operations and operations that create "impossible" situations are not possible, and allowing for verification of each step in the process in some qualitative or preferably quantitative way.

Removing human handling of the data, for example, in the form of keypunch input, helps eliminate clerical errors. This is not to say, obviously, that humans do not interact with the data, but that the interaction takes place on a "higher" level than specific numbers. Human interaction is most productive at the "object" level, i.e., entire parts, or easily conceptualized subsets of those parts. For example, cross-sections of the objects, or contours are suitable subsets of an object. Examples of ways of avoiding direct human input of numerical include automatic acquisition pf data via scanners, blue-print readers, analog-to-digital devices, and CAD software. where dimensioning and other specific numerical data tasks are done by computer.

Well designed software can, to some extent, aid in the elimination of errors by making certain operation illegal, or, at a more sophisticated level, check the results of operations to make sure that they do not create any "impossible" conditions. An example is the design of a medical prosthetic on a contour-by-contour basis; there must be some connectivity between adjacent contours or the cross-sections will not form a single object. The software must note if the designer has created this situation and inform the user of the immediate problem. "Impossible" conditions in the software or hardware, such as error conditions that might invalidate an operation, e.g., disk I/O errors, must also be detected when possible and reported to the user with an explanation of the possible impact on the processes in progress.

Verification of a design can involve many different criteria, for example, fit, stress analysis or some other measure. In designing medical implants, frequently the most important criterion is the physical goodness of fit. In the Cemax-1000, the fit is checked by graphically overlaying the implant on the original bone. Another good check of the fit is to generate a graphical image of the intersection of the implant and the original bone to check for overlapping or voids.

Verification of the correctness of the original data and subsequent operations on that data are naturally of great concern. At first glance, obtaining the data from a remote source using an automated digital scanner might obviate the need for initial data verification, but unfortunately, this is not the case. The data is obtained with a variety of different devices, any of which can be poorly maintained and operated. So checking of the new data is done as the data is reformatted into an internal standard data format. This format is the same regardless of the source of the data. Maintaining a common format eases the burden of programming and checking of certain simple parameters.

Once the data is inhouse, most of the responsibility for data integrity at the level of individual bits is on the hardware. Common approaches to maintaining data integrity at this low level include error detection and correction memory and check sums. As mentioned earlier, operations on the data at a higher level (e.g., thresholding and contouring) are checked visually by actual display of the data as objects. For implant modelling, the ultimate test of the process is actual experimental verification of fit.

III. Data Representation and Management

Representation

The two most popular ways to represent three-dimensional objects, boundary representation (B-rep) and solids geometry (CSG), both present a set of trade-offs from the programmers and users' points of view. For instance, using true solids modelling simplifies the data representation for objects that can be modelled as logical groups of certain primitives. Surface representations are more convenient when the objects modelled are of arbitrary complexity. We chose B-rep for this reason; human anatomy is not readily modelled using simple 3-D primitives. Alternative ways of representing the data, such as octtrees and complete voxel representations, are much more memory and computation intensive (Ref. 5 - 6). However, they do have the advantage of being algorithmically simplest and therefore easiset to implement in hardware. In the future, these two methods, especially the latter, will have increasing application.

Boundary representation itself encompasses a wide range of different methods, from surface patches for smooth objects to planar polygons for faceted approaches. Note that most boundary representations are converted into planar polygons before display to increase display speed. Contour files, or modelling the objects as a series of ordered cross-sectional slices, are the standard within Contour Medical Systems, since they are a natural representation for tomographic data; but have not historically been of wide spread use in the manufacturing engineering community. However, they do have the advantages of relatively compact storage and straight forward implementation.

Data structures are of fundamental importance, not only from the viewpoints of display and ease of modification, but from the programmers viewpoint, the simplicity of various algorithms and the efficiency of storage. Avoiding the obsolescence of the chosen data structures due to software or hardware advances, or changing requirements for the system is paramount. At CMS, the contour method of representation is universal, i.e., for display, modification and milling, and has so far proved adequate. There are potential serious limitations to this B-rep scheme; multi-object interference calculations, connectivity, and the forced homogeneity of the material. However, they do allow arbitrary precision, quick shading algorithms, and rapid interactive modification. The display and modification of the database are easily handled using available computer graphics algorithms. However, tool path generation generally proves to be a more difficult problem.

In order to maintain a high throughput of parts, a new method of tool path generation has been implemented which minimizes the difficulty for complicated 3-D parts. Instead of using software to calculate the tool path from basic geometric information, which is extremely difficult for arbitrarily shaped 3-D objects, the algorithm uses the computer graphics algorithm of visible surface generation. The user interactively selects the orientation of the part to obtain the view of the object which maximizes the amount of visible surface generation. The visible surface, the actual part of the object shown on the display, and the corresponding surface from a diametrically opposed point of view, are used to form two arrays of tool locations. The array represents the height of the tool above the surface (the intensity of the image at that point if depth-encoded shading has been performed). This is a general purpose algorithm which can be applied to objects with arbitrary resolution, by interpolation, with the chosen accuracy depending on the quality of finish desired.

Management

Data storage, retrieval and management issues can be important depending on the specification for the system. For instance, 10 megabyte files can be archived to tape in 20 minutes, and retrieved at the same rate with relatively low-cost streaming tape drives. If quicker access times are required very high volume magnetic or optical disks are needed. Large magnetic disk drives for long term shortage are not economical when the data throughput as on the order of 50 megabytes/CNC/day. Optical disk technology presently offers very large read-only storage systems. Read/write optical storage is in the not-too-distant future. An alternative to storing the entire files on magnetic disk is to store only summaries of the data. This data and data files with the production information can be automatically entered into the relational database system and used for the extraction of statistical, billing, and part tracking information.

One of the best methods of managing the revelant part data is relational database management software, which is available for a wide variety of hardware and operating systems, including super-micros. The software management functions should include modules for part tracking, report writing, database querying and user access to software tools for customizing use of the system. Relevant criteria for selection of database manager include expandability, speed, size of files allowed and flexibility. One other important feature is the generality of the software, it should be able to run on and communicate between a wide variety of different computers.

Relatively large amounts of data can be facilely handled after applying some of the ideas outlined above. As an example, at Contour Medical Systems we are presently using scan data files which range from 8 - 20 megabytes. This data is compressed to roughly a few hundred kilobytes by the process of thresholding and contouring; it expands slightly when converted to mill format which average 1/2 to 1 megabytes. All processing and data communication is electronic -- the traditional process of drawing blueprints, converting blueprints to papertape for the CNC, and then machining, would take much longer. For example, a typical time budget for a single part is as follows:

	Typical time in process	Typical size of file used in process
Read in 9-track tape	30 min.	10 megabytes
Threshold & contour	20 min.	10 megabytes
Generate & output surfaces	5 min.	0.3 megabytes
Smooth & make mill file	15 min.	0.5 megabytes
Mill both mold halves	40 min.	1.0 megabytes
Pouring	10 min.	
	2 hrs.	

Note however, these times represent computer run-times, not man-hours or even CPU hours. Since the system is multi-tasking, several processes can be run at once. The bottleneck, in fact, is the ability of the milling device to process the mill files and make the mold halves. The data throughput for a typical system must be as high as 40 - 50 megabytes for an 8-hour day for economical operation.

Conclusion

A relatively low cost super-micro based system was implemented to achieve the completely "paperless" design and machining of custom therapeutic parts. A large degree of automation of the entire process was necessary to obtain high throughput of data. The description of arbitrary 3-D parts uses a boundary representation scheme based on contours of the desired object. Contours are a universal data structure, independent of their origin whether from thresholded data or obtained from interactive input by the user, which lend themselves to a relatively compact representation and to efficient display, design, and milling algorithms.

The special nature of the data acquisition, i.e., via automated scanners eliminates the burden of manual data entry. However, the same issues of data verification and management apply to this environment, and in fact, become more critical due to the reliance on automatic processing and the speedier propagation of errors. Maximization of computer automation as applied to the production of "quantity one" cast parts has resulted in a financially feasible turnaround time of two hours.

References

Dev, P., Wood, S., Duncan, J.P., and White, D.N., "An interactive graphics system for planning reconstructive surgery", Proc. of the National COmputer Graphics Assoc., Chicago, IL, pp 130 - 135. June 1983.

Rhodes, M.L., et al, "Anatomic model and prosthesis manufacturing using CT images", National Computer Graphics Proceedings, v, III., pp 110 - 124, 1985.

Vannier, M.W., "Surface reconstruction from CT scans", Surgical Rounds, pp. 20 - 27, March 1984.

Hemmy, D.C., David, D.J., and Herman, G.T., "Three-dimensional reconstruction of craniofacial deformity using computed tomography", Neurosurgery, pp. 13:534 - 541, 1983.

Meagher, D.J., "Geometric modelling using octree encoding", Computer Graphics and Image Processing, v. 19, 1982.

Goldwasser, S.M. and Reynolds, R.A., "An architecture for the real-time display three-dimensional objects", Proceedings of the 1983 International Conference on Parallel Processing, pp 269 - 274, 1983.

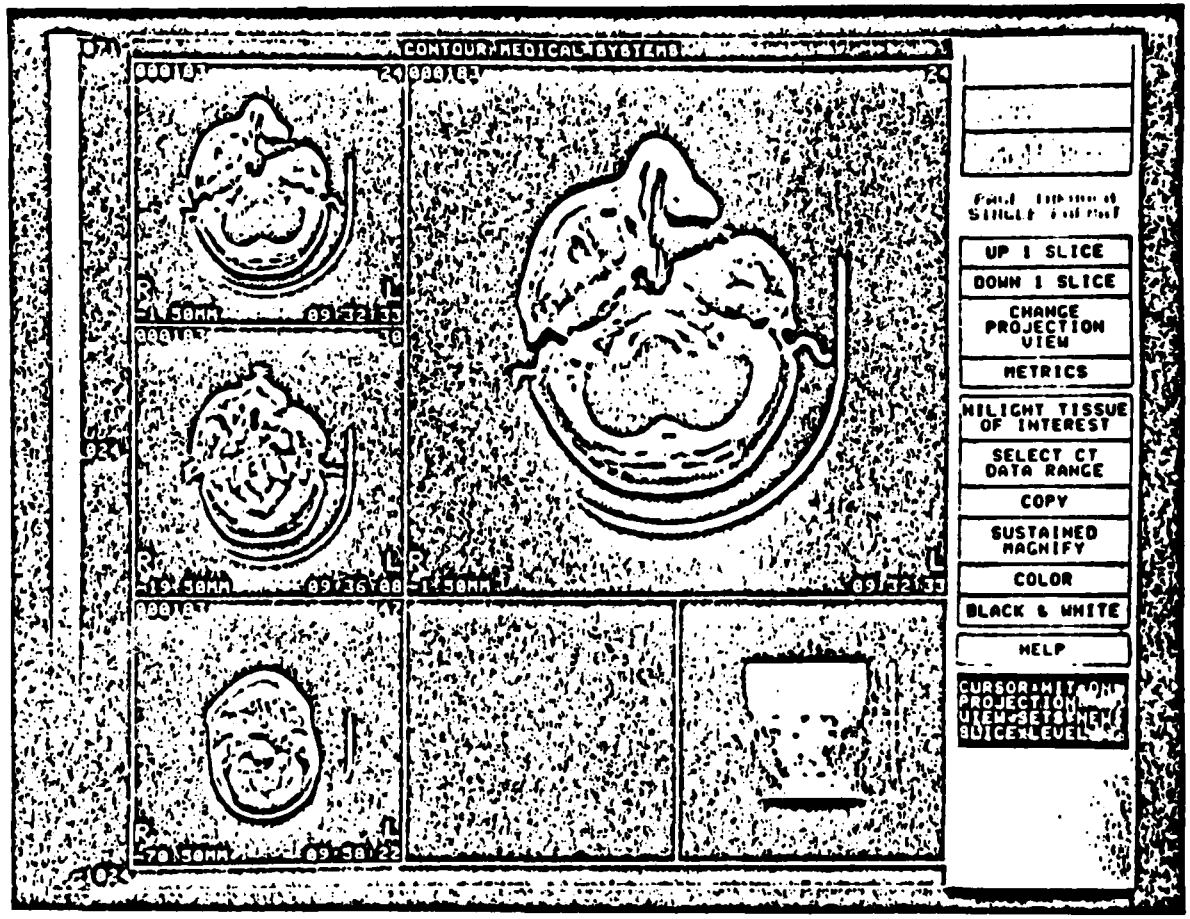

CONTOUR MEDICAL SYSTEMS
UP 1 SLICE
DOWN 1 SLICE
CHANGE PROJECTION VIEW
METRICS
HILIGHT TISSUE OF INTEREST
SELECT CT DATA RANGE
COPY
SUSTAINED MAGNIFY
COLOR
BLACK & WHITE
HELP

Figure 1

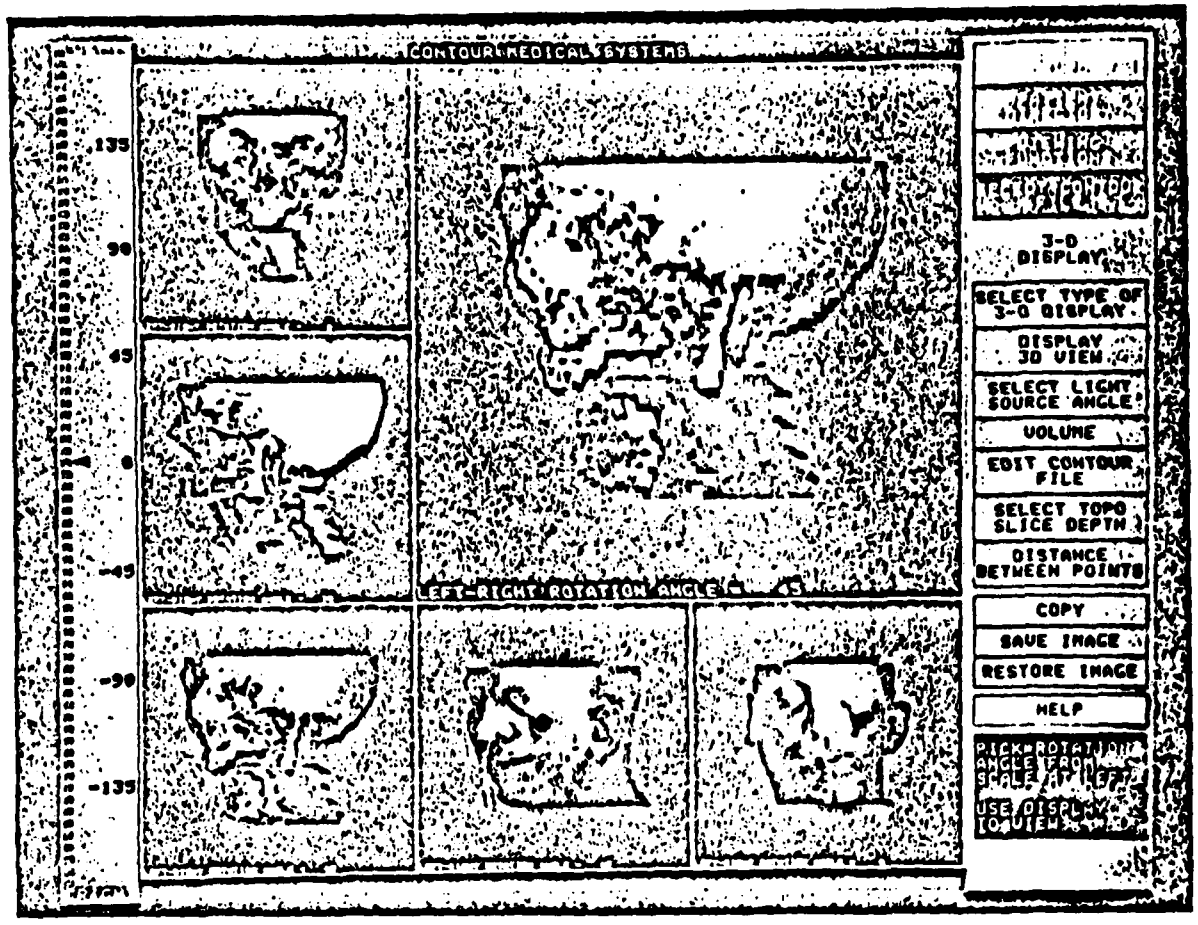

CONTOUR MEDICAL SYSTEMS
135
90
45
0
-45
-90
-135
LEFT-RIGHT ROTATION ANGLE
3-D DISPLAY
SELECT TYPE OF 3-D DISPLAY
DISPLAY 3D VIEW
SELECT LIGHT SOURCE ANGLE
VOLUME
EDIT CONTOUR FILE
SELECT TOPO SLICE DEPTH
DISTANCE BETWEEN POINTS
COPY
SAVE IMAGE
RESTORE IMAGE
HELP
USE DISPLAY TO VIEW

Figure 2

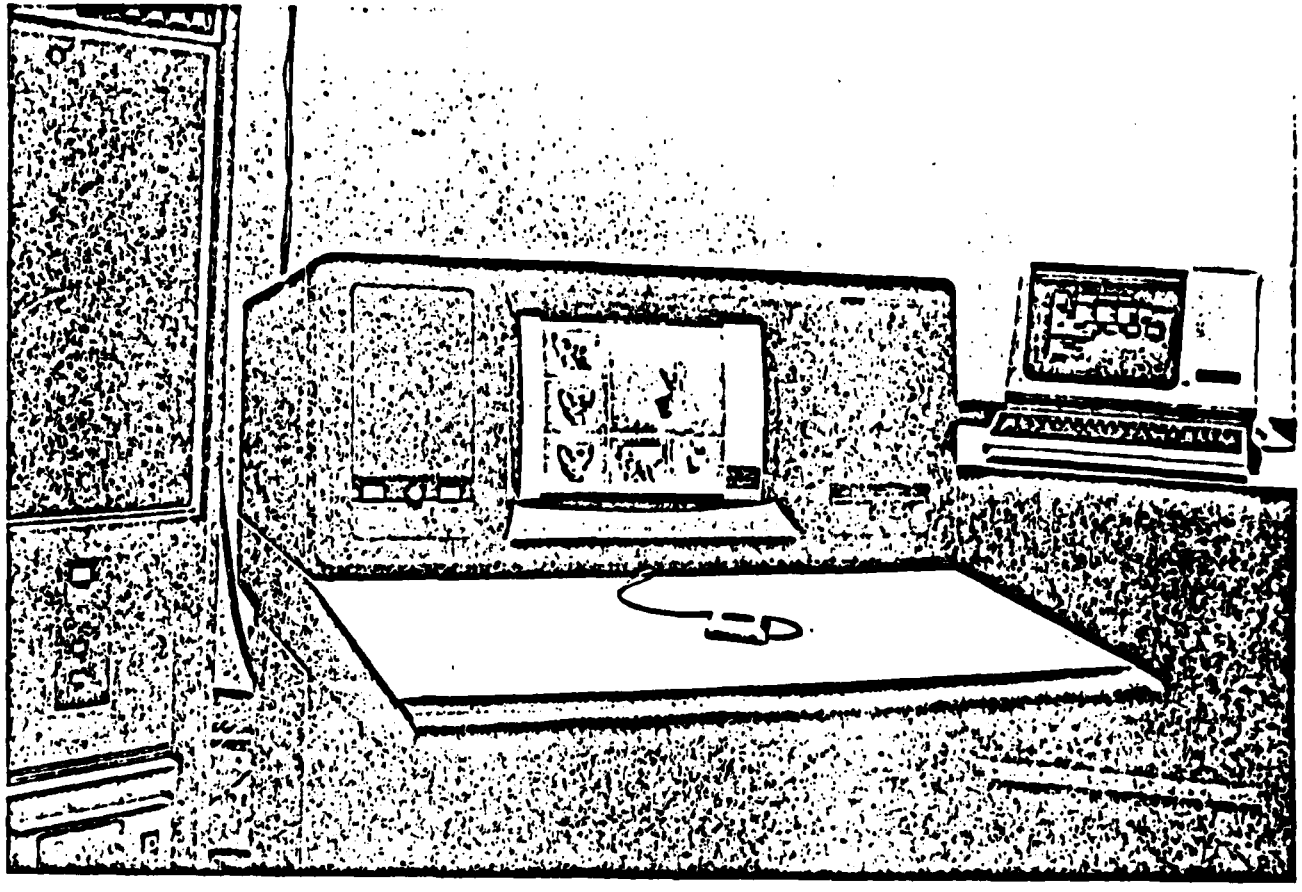


Figure 3

International Workshop on Physics and Engineering in Computerized Multidimensional Imaging and Processing

April 2 - 4, 1986

at the

University of California, Irvine

in cooperation with

Department of Radiological Sciences
University of California, Irvine,
University of California, Irvine Extension

A three-day workshop on ***Physics and Engineering of Computerized Multidimensional Imaging and Processing will*** convene at the University of California, Irvine between April 2 - 4, 1986.

Since the inception of these workshops, the first one held in Irvine in 1979 and the subsequent one at Pacific Grove in 1982, two important areas of imaging have progressed substantially, namely the emission computed tomography (ECT) and nuclear magnetic resonance (NMR) imaging. In the ECT area, progress in single photon tomography is so great that this modality is currently accepted as an important clinical tool. The spatial resolution capability of positron emission tomography is now approaching the ultimate limitation imposed by finite positron range and emission angular uncertainty. More recently NMR imaging has become one of the most significant advancements in biomedical imaging research and commercial development. It is now believed that industrial and governmental agencies have already spent over a billion dollars on the NMR related research and commercial development. NMR imaging is now accepted not only as a clinically useful imaging modality but also as a new diagnostic technique which will naturally expand into many other areas of physical and biomedical research such as the study of fundamental metabolism of living organs by in-vivo observations of the phosphor kinetics. The advancement of this technique which is sometimes known as the 4-D NMR imaging poses a challenge to create yet unknown NMR imaging methods.

Although the main emphasis of the meeting will be in the area of medical imaging, contributions in other non-medical areas are strongly encouraged. Indeed, the techniques and methods developed for medical imaging could, in principle, be applied to other areas of physical and biological sciences with appropriate modifications. Some of the most recent non-medical applications being in the area of synchrotron radiation microscopic tomography and electron spin resonance (ESR) tomography. In this meeting, we would like to solicit papers and ideas on multidimensional imaging techniques, associated signal processing methods, computational strategies for 2-D, 3-D and 4-D image signal processing and new application areas. Throughout the meeting, we will not only attempt to scrutinize the established techniques but also seriously search for new directions.

We believe that the previous two meetings held here have set the tradition to review the field with great depth and breadth. As such, the next meeting to be held in 1986 is expected to be useful for the experts and in addition set the future direction for the field by cross fertilization of ideas.

It is interesting to remember the trends of the two previous meetings that were held. In the first meeting, the major emphasis was on the image reconstruction algorithms, X-ray CT, emission CT, and ultrasound. NMR was discussed by Dr. Lauterbur from SUNY-Stony Brook. In the second meeting, emission CT, especially various PET devices, dynamic X-ray CT, and NMR were the dominant topics. In the second meeting, however, NMR was clearly identified as a new and definitely useful diagnostic modality.

In the forthcoming meeting, we predict that NMR will be a major topic, while emission CT, both PET and SPECT will also be treated as active, important areas. We hope that some new unknown exotic techniques and ideas would also emerge at the meeting. The last goal could only be achieved through the active participation of the attendees and by interaction of other related scientific fields.

Contributions are encouraged from interested authors who are planning to attend the Workshop. Each author should submit 3 copies each of an abstract and a 500-word summary. The abstracts must be on a separate sheet. The summary will be used as a basis for paper selection for presentation and publication in the proceedings.

Z.H. Cho, Ph.D.
O. Nalcioglu, Ph.D.
T.F. Budinger, M.D., Ph.D.

Abstract and Summary Deadline: Received by November 1, 1985

Send three (3) copies of each to:

O. Nalcioglu, Ph.D.
Department of Radiological Sciences
University of California, Irvine
Irvine, CA 92717, USA
Telephone: (714) 856-5904

17 May 1985, Volume 228, Number 4701 SCIENCE

4.4.1

Positron Emission Tomography: Human Brain Function and Biochemistry

Michael E. Phelps and John C. Mazziotta

The study of human brain function and its alterations with disease remains one of the most challenging and intriguing scientific issues of our time. Advances in brain research are increasing our understanding of the biochemical nature of the brain and are demonstrating that the earliest and most specific changes occurring in diseases of the brain are those that disturb its underlying biochemical processes. Until recently, we have had little direct access to the local biochemistry of the living human brain. Inferences about the chemical status of the brain are typically made by chemical assays of blood, cerebrospinal fluid, and urine or occasionally through biopsy procedures. With the development of positron-emission tomography or PET (*1*) these disturbances, along with the study of normal cerebral function can be investigated in humans (*2, 3*).

Summary. Positron emission tomography (PET) is an analytical imaging technique that provides a way of making in vivo measurements of the anatomical distribution and rates of specific biochemical reactions. This ability of PET to measure and image dynamic biochemistry builds a bridge between the basic and clinical neurosciences founded on the commonality of the types of measurements made. Clinical findings with PET in humans are suggesting hypotheses that can be tested rigorously in the basic science laboratory.

Knowledge concerning the biochemical basis of human disease should aid in developing earlier (when containment or reversibility of disease is more probable), more specific, and improved therapies. Therapeutic interventions for human brain disorders attempt to remove, block, or supplement chemical processes of the brain that have been altered by disease. For example, surgical therapies can resupply nutrients to deprived tissue by revascularization or can remove altered processes by tissue resection. Drug therapies are targeted at chemical reaction sequences that have been perturbed by disease, resulting in altered mood states and neurological dysfunction. In addition, local determinations of biochemical processes could provide more objective evaluations of therapeutic responses, better prognostic indicators, and improved differential diagnoses of diseases that are clinically homogeneous yet have diverse chemical alterations.

Equally important to the study of human diseases of the brain is the need to develop a better understanding of the structure, organization, and chemical basis of normal cerebral function. Although there is a rapid growth of knowledge occurring in this area from animal, isolated cell, and biochemical environments, corresponding direct investigations of the living human subject are needed. This, in part, results from the fact that there are anatomical areas and physiological functions found only in the human brain. In addition, identifying and treating diseases of the human brain require a better understanding of its normal processes. New scientific techniques such as PET provide one means to combine basic and clinical research to achieve these goals.

The PET Method

The use of PET to obtain quantitative biochemical rates in vivo requires the integration of three major components: compounds labeled with radioisotopes, the positron tomograph, and tracer kinetic mathematical models. These have each been reviewed (*2, 3*) but are briefly discussed below.

Labeled compounds. One of the attractive aspects of the PET technique is that the compounds of interest can be labeled with radioisotopes of natural elements of the biochemical constituents of the body. For example, natural isotopes of carbon, nitrogen, and oxygen are replaced with the short-lived radioisotopes carbon-11, nitrogen-13, and oxygen-15. Fluorine-18 is used as a substitute for hydrogen. These isotopes all decay by the emission of positrons (antielectrons) that combine with electrons to produce two 511-kiloelectron volt gamma rays, which are emitted 180° apart, which easily penetrate the head, and which then allow external detection. The only radioisotopes of these elements that can be detected outside the body are positron emitters. Because carbon, nitrogen, and oxygen are the constituents of virtually all biomolecules and drugs, in principle, an unlimited number of biologically active substrates can be labeled with these radioisotopes without disrupting their biochemical properties (*3, 4*).

Various chemical and biosynthetic procedures have been used to label more than 200 biological substrates and drugs with these isotopes; these labeled compounds constitute a large potential resource for development of bioassay methods with PET (*4*) (Table 1). This list includes labeled amino acids, carboxylic

Michael E. Phelps is Jennifer Jones Simon Professor and Chief, Division of Nuclear Medicine and Biophysics, Department of Radiological Sciences and John C. Mazziotta is an assistant professor, Department of Neurology and Radiological Sciences, UCLA School of Medicine and the Laboratory of Nuclear Medicine Los Angeles, California 90024.

acids, amides, amines, nitriles, alcohols, sugars, hydantoins, steroids, and their derivatives as well as specific substrates, metabolites, analogs and drugs. Rapid semiautomated techniques have been and continue to be developed to meet the needs of PET, as required by the short half-lives of these isotopes (2 to 110 minutes). Although we discuss only compounds labeled with ^{11}C, ^{13}N, ^{15}O, and ^{18}F, there are also positron-emitting isotopes of Rb, Fe, Mn, Na, K, P, Br, Kr, I, and others.

Positron tomograph. The tomograph consists of an array of radiation detectors that are placed circumferentially around the head and record the emission of γ-rays from the tissue distribution of positron activity. Data collected in this manner are used to form a tomographic image of the cross-sectional distribution of tissue concentration of radioactivity according to the principles of computed tomography (*1*). This provides a quantitative, noninvasive measurement in humans analogous to the well-known invasive techniques of quantitative autoradiography or external counting of resected tissues samples in animal studies. In the invasive techniques, ^{14}C and ^{3}H are commonly used, whereas in PET ^{11}C, ^{13}N, ^{15}O, and ^{18}F are used. This quantitative tissue assay capability of PET provides the means for implementing tracer kinetic methods used throughout the basic sciences.

Tracer kinetic models. The third major component of PET brings together principles of labeled compounds that trace hemodynamic, transport, and biochemical processes, as well as the tissue radioassay capabilities of the tomograph with mathematical models of reaction sequences to provide a framework for calculation of the rates of processes under study. These models, when applied to labeled compounds (tracers) are called tracer kinetic models. The models represent mathematical descriptions of transport or biochemical reaction sequences. Each segment of the sequence is described as a "compartment," and differential equations describe the movement of the natural substrate or labeled compounds (or both) between these compartments. For example, A ⇄ B ⇄ C represents a reaction sequence where A, B, and C are compartments. Measurement is made of the flux between compartments which in turn is used to determine the rate at which the reaction sequence proceeds. These compartments can be separated by membranes where facilitated, active, or passive diffusion may occur or may represent the separation of chemical reactants and products. The configurations for these compartmental models are obtained from knowledge of hemodynamic, transport, and biochemical systems.

For example, a simple two-compartmental model is used in the measurement of oxygen metabolism with $^{15}O_2$. The first compartment is the plasma and the second is tissue where the oxygen is metabolized to water. The rate of transport from the first compartment to the second is described by a rate equation from which the rate of tissue extraction and metabolism of oxygen can be determined. As a reaction sequence becomes more complex, the number of compartments will increase. Some principles and aspects of tracer kinetic methods and their use in PET are as follows:

1) The tracer (that is, the labeled compound) is very low in mass compared to the compound being traced, such that there are no significant mass effects that would produce physiologic perturbations altering the system under study. The low mass of these tracers allows PET measurements of reaction rates of substrates with concentrations of less than a few picomoles per gram. However, one can also increase the mass of the tracer compound to make measurements under physiologic loads (for example, drug effects).

2) When the rate of the reaction being studied is not changing during the measurement (steady state), the net reaction rate of any one step in a nonbranching sequence is equal to the net rate of the whole reaction sequence. Thus, measurement of the net initial rates of a reaction sequence with a labeled substrate can be used to assess the net flux of the entire pathway. Alternatively, substrate analogs that isolate one or a small number of steps in a complex reaction sequence can be used to measure the net reaction rate of the whole sequence. Since the analog may have somewhat different reaction kinetics from that of the natural substrate, correction terms based on the principles of competitive reactant or substrate kinetics are used to transform the measured rate of the analog into the corresponding value of the natural compound. These are fundamental principles of the fields of biochemistry and pharmacology. The deoxyglucose model of Sokoloff *et al.* (*5*) is an example of this approach, and under the assumptions of the model has been shown to provide accurate estimates of cerebral glucose utilization rates. Except for Fig. 6b all data on rates of glucose utilization in this article were obtained with the use of 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG), which, like 2-deoxy-D-glucose (DG) (*5*), has been shown to be an excellent substrate for hexokinase, and like DG, does not undergo further reactions in the glycolytic pathway during the time course of the PET measurements (*6–8*).

3) Although measurements of chemical substrate concentrations are provided by kinetic tracer methods, their use for determinations of chemical reaction rates is more important. Rates of a reaction can change without changes in sub-

Table 1. Partial list of compounds labeled with positron-emitting radionuclides.

Cerebral blood flow
- $H_2{}^{15}O$, $C^{15}O_2$*, ^{77}Kr, $CH_3{}^{18}F$, ^{18}F-labeled antipyrine, [^{11}C]alcohols, ^{18}F-labeled ethanol

Cerebral blood volume
- ^{11}CO, $C^{15}O$, ^{68}Ga-labeled EDTA†

Cerebral tissue *p*H
- [^{11}C]DMO, $^{11}CO_2$

Transport and metabolism
- Oxygen
  - $^{15}O_2$
- Glucose, glucose analogs, and metabolites
  - 2-deoxy-2-[^{18}F]fluoro-D-glucose, 2-[^{11}C]deoxy-D-glucose, [^{11}C]D-glucose, 3-*O*-[^{11}C]methyl-D-glucose, [^{11}C]lactate, -pyruvate, -acetate, -succinate, -oxaloacetate
- Amino acids: ^{13}N-labeled
  - L-[^{13}N]glutamate, -α and ω-glutamine, -alanine, -aspartate, -leucine, -valine, -isoleucine, -methionine
- Amino acids: ^{11}C-labeled
  - L-[^{11}C]aspartate, -glutamate, -valine; D,L-[^{11}C]alanine, -leucine, -tryptophan, -1-aminocyclopentane carboxylic acid, -1-aminocyclobutane carboxylic acid
- Free Fatty Acids
  - [^{11}C]palmitic acid, -oleic acid, -heptadecanoic acid, -β-methylheptadecanoic acid

Molecular diffusion
- ^{68}Ga-labeled EDTA, ^{82}Rb

Protein synthesis
- L-[1-^{11}C]leucine, -methionine, -phenylalanine, L[^{11}C-*methyl*]methionine

Receptor systems
- Dopaminergic
  - [^{18}F]spiperone, [^{11}C]spiperone, [^{75}Br] and [^{76}Br]-*p*-bromospiperone, [^{18}F]haloperidol, [^{11}C]pimozide, [^{11}C]methylspiperone, L-[^{11}C]dopa, [6-^{18}F]-fluoro-L-dopa
- Cholinergic
  - [^{11}C]imipramine, [^{11}C]QNB†
- Benzodiazepine
  - [^{11}C]flunitrazepam, [^{11}C]diazepam, [^{18}F]fluoro valium
- Opiate
  - [^{11}C]etorphine, *N*-methyl-[^{11}C]morphine, -heroin, -carfentanil
- Adrenergic
  - [^{11}C]norepinephrine, [^{11}C]propanolol

Anticonvulsants
- [^{11}C]valproate, [^{11}C]diphenylhydantoin

*Converted to $H_2{}^{15}O$ in lungs after inhalation. †QNB, quinuclidinyl benzilate

strate concentration in the open system of the brain (and the whole body). Changes in tissue substrate concentration are also not specific to the magnitude or direction of changes in rates of reactions.

4) While the chemical form to which the label is attached is known at the time it is intravenously injected or inhaled, the positron tomograph only measures the kinetic changes (changes in time) of tissue concentrations of the label spatially throughout the brain. In order to convert these images of tissue radioactivity concentration to measurements of local reaction rates in vivo, PET measurements must be combined with the time course of the unlabeled or labeled substrate in blood and properly formulated and validated tracer kinetic models. (For example, in determinations of blood flow and volume and of drug interactions only the tracer concentration is required.) With measurements carried out in this manner, the tomograph can provide accurate data on local reaction rates in man.

5) Because of the short half-lives of positron-emitting isotopes it is possible to perform multiple studies in a single setting to observe changes in spontaneous or stimulus-induced alterations in behavior or, using different tracers, for measuring different biochemical processes. For example, ^{15}O measurements can be made in times on the order of 30 seconds and repeated at about 8-minute intervals.

Quantitative PET Methods: Present Status

In autoradiographic studies with ^{14}C-labeled DG, the experiment is terminated 40 minutes after injection of DG. However, in PET with FDG, scans are started and completed within time frames ranging typically from 40 to 100 minutes after injection. Measurements over this time are needed because of (i) the time required to collect enough counts to form the tomographic images of the whole brain, (ii) delays or changes in the study due to problems associated with patients, or (iii) varying time requirements of different types of study protocols (*2, 3*). Because of the longer times between injection and measurement, slow dephosphorylation of FDG 6-phosphate, which is not a usual problem in autoradiography, has been shown to occur and has been taken into account by extending Sokoloff's original model (*5*) to include this reaction (*7*). This restores the accuracy of the model calculation of glucose utilization rates with FDG at these later times and also provides a method for investigating this reaction pathway (*7, 9*).

The lumped constant (LC) in the deoxyglucose model is a term based on the principles of competitive substrate kinetics and accounts for the differences between glucose and deoxyglucose affinities for the transport carrier system and hexokinase. The value of LC has been shown to be regionally invariant in the brain but of a different magnitude in different species, thus appropriate values for the species under study must be used (*5*). Although LC has been shown to be stable during various states such as changes in blood flow and anesthesia induced hypometabolism (*5*), and in chronic human ischemia (*9*), it predictably changes when the rate limiting step shifts from phosphorylation to transport such as in severe hypoglycemia (*10*) and status epilepticus (*10a*). In these cases LC will increase, and the resultant calculated values of glucose utilization will be subject to considerable error if this effect is not taken into account. Anatomical localization and identification of changes in glucose utilization are still correct, but increases and decreases will be overestimated. Methods are being developed for measuring local values of LC to determine when such conditions are encountered and the magnitude of the effect (*11*). The DG and FDG method have also shown excellent agreement with the "gold standard" for measurement of glucose utilization, the Kety/Schmidt method (*12*) that includes measurement of the product of blood flow and the arteriovenous difference for glucose across the brain (*3, 7, 8, 13*).

Fig. 1. Structural compared to functional anatomy. Images obtained with three different methods from a patient with multiple infarct dementia. Patient had x-ray CT (center row) and PET studies of glucose utilization with FDG (bottom row) on the same day. Seven days later, the patient died of nonneurological causes; gross and microscopic evaluations of the brain (top row) were then made. The two forms of structural imaging (x-ray CT and postmortem) and the metabolic study with PET both demonstrated multiple small infarctions (arrows) of deep structures of the brain (striatum, putamen, thalamus, and internal capsule). Neither structural imaging techniques demonstrated abnormalities of the cortex. PET, however, demonstrated widespread abnormalities of frontal cortex (glucose utilization decreased 21 percent relative to contralateral side) particularly on the left (arrowheads). These distant effects probably represent disruption of afferent and efferent fiber systems between the frontal cortical areas and subcortical zones, most likely resulting from small subcortical infarcts seen structurally and metabolically. The global rate of glucose utilization was 0.19 $\mu mol\ min^{-1}\ g^{-1}$, about 30 percent lower than normal age-matched controls. [Courtesy of E. J. Metter, with J. C. Mazziotta and M. E. Phelps, UCLA School of Medicine]

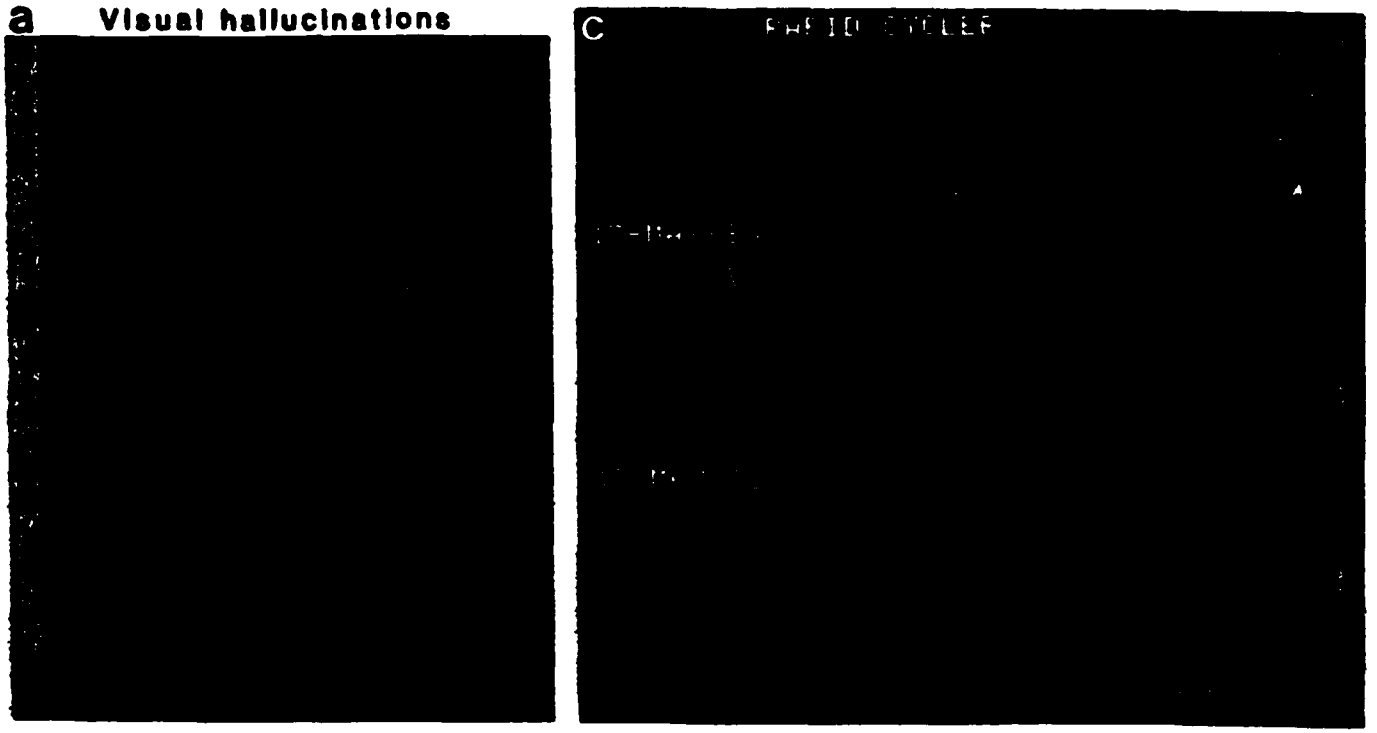
Models based on the principles of tracer kinetics have also been developed for autoradiography and PET with the use of ^{14}C-labeled (*14*) and ^{11}C-labeled (*15*) D-glucose. Although these models are based on the use of natural labeled substrate and therefore have the advantage, compared to DG, that LC = 1, by necessity they also contain approximations and correction terms. The most difficult of these is the release of labeled products of metabolism (such as CO_2 and lactate) from the tissue, an effect that must be taken into account by correction terms in the model. To minimize these problems, measurements must be made at early times after injection where errors are larger because of a greater dependence of model parameters on blood flow, blood volume, and the exact values of the rate constants for transport and metabolism of glucose. The retention of the metabolic product (DG-6-phosphate) with DG is advantageous in this respect since it is retained in tissue with a very slow clearance rate. The problems with labeled natural glucose are less difficult with PET because kinetic studies can be performed to measure the variables in the model directly (rather than to use assumed average values) in each experiment. Further studies, nevertheless, are necessary to better understand the magnitude of these issues, and to formulate better models and protocols, and to investigate the value of using glucose la-

beled at selected positions (such as, selective labeling at the carbon -1, -2, -4, or -6 positions).

Quantitative PET methods for oxygen transport and metabolism, blood flow, blood volume, glucose transport, and utilization have been widely used in PET programs around the world (*2, 3*). Intensive efforts are being focused on the study of tracer kinetic models for determining rates of protein synthesis, amino acid metabolism, tissue *p*H, molecular diffusion through the altered blood brain barrier, and neurotransmitter or receptor interactions with labeled substrates and ligands (Table 1). The results of these projects should provide the basis for some of the PET methods of the future (*3*).

The measurement of biochemical reaction rates with PET requires (i) adherence to the criteria used to develop and structure the tracer kinetic models, and (ii) continuing investigative efforts to better define the accuracy of these methods in all the states of cerebral function and dysfunction to which they will be applied. These requirements are necessary if measurements by different groups of investigators are to be compared. In this regard, PET is no different from any other biochemical measurement where the principles of enzyme and chemical reaction kinetics must be strictly followed.

There are limitations intrinsic to tracer kinetic model approaches along with uncertainties in spatial, temporal, and statistical factors from the tomograph, and we must recognize that ambiguities can arise from the present limited understanding of the biochemical nature of the brain. However, the ability to make such measurements in the living human brain more than compensates for the extensive developmental work which is required in the validation of a new PET method.



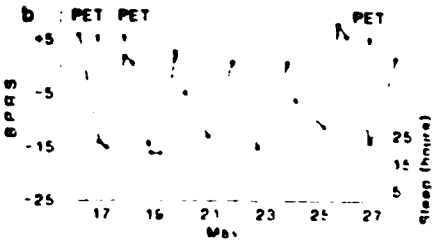


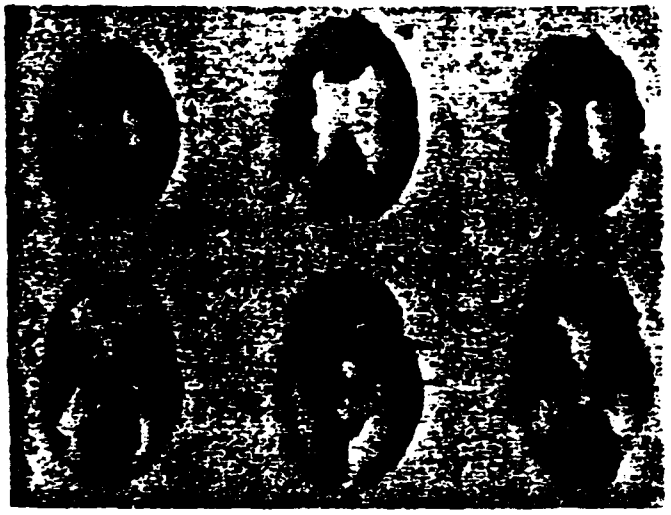
Fig. 2. Pathophysiology of changing behavioral states shown with PET. (a) Patient with complex partial epilepsy studied during seizures (ictal), after seizures (postictal), and after successful therapy (normal). In the ictal state (A) seizures began clinically with formed visual images in left superior quadrant of visual field followed by loss of consciousness and automatisms. Ictal EEG demonstrated discharges in right occipital lobe that quickly spread to the ipsilateral temporal lobe. PET demonstrated increased (284 percent relative to interictal state) glucose utilization in the right occipital and posterior temporal lobes with profound decreases in remainder of brain (42 percent relative to the interictal state). One month after a series of seizures, the patient, in a repeat study showed low glucose utilization (B) in the right occipital and posterior temporal lobes (one third of ictal values) where, during seizure, it was high. The low glucose utilization in the visual cortex (arrows) during this time is consistent with the patient's loss of vision in left visual field (homonymous hemianopia). After drug therapy, the patient remained seizure-free for 1 year. The glucose utilization during this time returned to normal (C), consistent with the resolution of the left visual field deficit. These studies show that PET can be used to detect both positive (hallucinations) and negative (visual field deficit) clinical manifestations of an epileptic disorder and the reversal of these changes resulting from treatment and clinical resolution of symptoms. The x-ray CT study was normal and unchanged in each state. [From Engel *et al.* (*18*); courtesy of *Neurology*] (b) Graph of the mood fluctuations of a rapid cycling bipolar (manic-depressive) affective disorder patient. The patient cycled between mood states every 24 to 48 hours. The modified brief psychiatric rating scale on the left (B.P.R.S.) indicates relative mania (positive values) and relative depression (negative values). Sleep cycles varied in parallel with mood swings. Three studies of cerebral glucose utilization with PET were obtained, two during depressive states and one during hypomania. [From Baxter *et al.*, in (*22*); courtesy of *Archives of General Psychiatry*] (c) Glucose utilization images of the rapid cycling bipolar patient described in (b). The scale to the right is glucose utilization rates in micromole per minute per 100 grams. Studies on 17 May 1983 and 27 May 1983 were made when the patient was in depressed state; the study on 18 May 1983 was made during a hypomanic state. Global reductions in glucose utilization (relative to the scan obtained during hypomania) can be seen for the two studies obtained in the depressed state. The global supratentorial increase in glucose utilization from depression to hypomania was about 40 percent. The global glucose utilization rate in the hypomanic state is not significantly different from that of age-matched normals. These studies demonstrate the ability of PET to provide pathophysiological insights into abnormal behavior that occurs as either a manifestation of epilepsy or psychiatric disorders. [From Baxter *et al.*, in (*22*), courtesy of *Archives of General Psychiatry*]

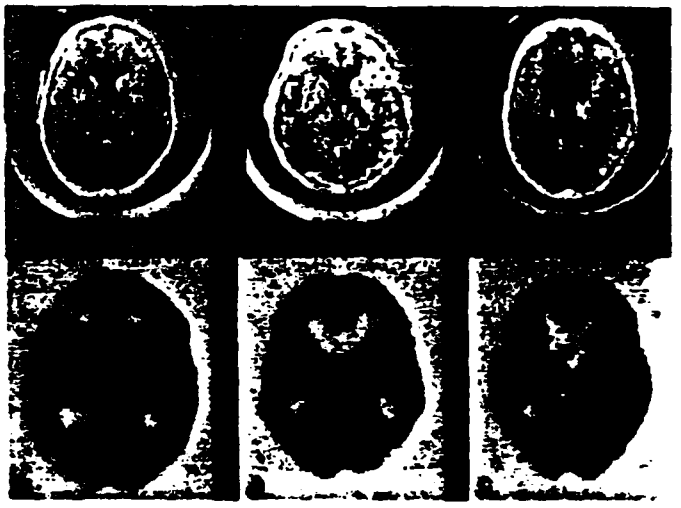
Cerebral organization and structure-function relationships. Methods used to investigate the structural nature of the human brain include in vivo techniques, such as x-ray computed tomography (CT) and proton nuclear magnetic resonance (NMR) CT, and postmortem techniques. Relatively little information exists about local biochemical and physiological processes of the human brain in vivo. Knowledge of the functional organization of the human brain has, however, been obtained through PET studies of both normal individuals and patients with cerebral disorders. Such studies provide a more comprehensive view of cerebral organization than structural studies alone (Fig. 1), particularly when one attempts to correlate the resulting data with behavioral observations of the subject.

Since all diseases of the brain result from or produce biochemical alterations, functional images that display such processes have demonstrated earlier, larger, and more distributed lesions than those found in anatomically oriented techniques including detailed postmortem evaluations (Fig. 1). Examples of the mismatch between structural and functional lesions have been demonstrated in patients with seizure disorders (Fig. 2a), dementing processes (Fig. 3), neurodegenerative diseases (Fig. 3), and acute cerebral infarcts (see Fig. 5).

Since the brain consists of large numbers of interconnected substructures, damage to one structure or its interconnecting fiber bundles will also result in functional effects at multiple sites throughout any given network. PET has revealed this distributed organization, leading to a more comprehensive view of human functional brain systems in health and disease. Traditional clinical-pathological correlations that have been the mainstay of symptom-lesion localization in the brain may soon give way to "clinical-physiological correlations" (*16*) that can be performed in vivo with PET.

Structural evidence has been scant in the search for the basis of a number of human cerebral disorders. Examples of entities in this category would include psychiatric syndromes, many forms of





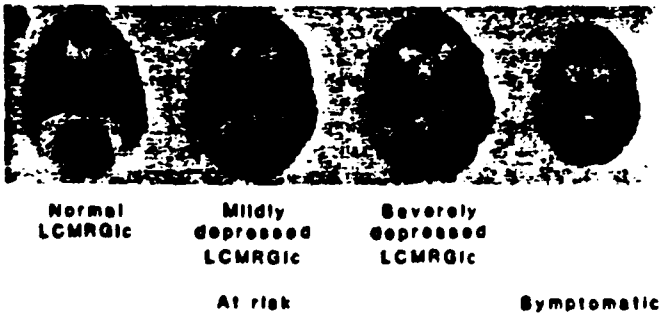


Fig. 3 (a) Differential diagnosis and pathophysiology with PET. Three elderly patients with abnormal glucose utilization. Column D images are from a patient with chronic depression (pseudodementia) who demonstrated decreased glucose utilization (left to right 0.91 ± 0.08 compared to 1.00 ± 0.05 in controls) of the left inferior frontal cortex (arrows) but was otherwise normal. Images in the column labeled MID are from a patient with multiple infarct dementia demonstrating multiple focal areas of cortical and subcortical decreased glucose utilization (from 17 to 48 percent) compared to global values (arrows) resulting from direct and remote effects of small cerebral infarctions. The column labeled AD is from a patient with moderately severe Alzheimer's disease. This patient shows extremely low glucose utilization of the posterior parietal (47 percent decrease) (upper image) inferior frontal (28 percent decrease) and temporal cortex (25 percent decrease) (bottom image) with relative sparing of the primary visual, sensory-motor cortical and subcortical zones (about 12 percent decrease). Alzheimer's disease patients have also shown an average decrease of global supratentorial glucose utilization (33 percent), compared to age-matched normal control subjects, with the major abnormalities occurring in neocortical zones. This study demonstrates the differential diagnostic capability of PET to separate patients with processes affecting mental abilities based on pathophysiological patterns observed in the functional images. Stated values are for groups of patients. Errors are standard deviations. [From D. E. Kuhl *et al.* (*47*), courtesy of *Radiology*.] (b) Functional abnormalities in Huntington's disease. x-ray CT (top row) and local cerebral glucose utilization with PET (bottom row) in Huntington's disease. Left column is from a normal subject and demonstrates the normal structure and glucose utilization of the caudate nucleus (arrows). Patient in center column has early clinical symptoms of Huntington's disease and demonstrates a normal structural appearance of the caudate nucleus on x-ray CT image but decreased glucose utilization in caudate and adjacent basal ganglia structures (putamen and globus pallidus). The column at the right is a patient with late Huntington's disease and demonstrates both structural (cortical and subcortical atrophy) and functional abnormalities of the caudate and putamen bilaterally. [From D. E. Kuhl *et al.* (*31*); courtesy of *Annals of Neurology*.] (c) Subjects at-risk for Huntington's disease. Image at far right is from a patient with symptomatic Huntington's disease demonstrating loss of glucose utilization of the basal ganglia as was seen in (b). The three images on the left are all from asymptomatic at-risk subjects (offspring of Huntington's disease patients) each having a 50 percent chance of developing the disorder. In this group (15 subjects) about half of the patients have demonstrated a normal glucose utilization pattern while the other half show mild to severely depressed glucose utilization in the caudate. The at-risk subjects were symptom free at the time of the study. This study indicates that PET may serve to identify physiological abnormalities that not only precede structural changes in the brain but also preceded the onset of symptoms in susceptible subjects. [From D. E. Kuhl *et al.* (*31*), courtesy of *Annals of Neurology*.]

epilepsy), and various developmental disorders of the brain. In patients with partial seizure disorders (symptoms referable to a limited region of the brain) interictal studies (between seizures) with FDG have identified areas of decreased glucose utilization in 70 percent of affected individuals (Fig. 2a) (*17*). These zones could be correlated with the site of maximal abnormalities determined by surface and depth electrode electrophysiological techniques, and were found to be extremely specific in identifying sites of microscopic pathology not detected by the conventional radiological imaging techniques such as x-ray CT, angiography, and pneumo-encephalography (*17*). During seizure activity (ictal period), sites that show low glucose utilization interictally show increased utilization (often increasing by 100 to 200 percent) (Fig. 2a); this suggests that the interictal low glucose utilization is at least in part due to nonstructural causes (*17, 18*). The sites of increased glucose utilization during the ictal phases of partial seizures correlated both with the behavior of the patient and with spikes recorded from scalp and particularly depth electroencephalographic (EEG) recordings (Fig. 2a) (*17, 18*). Patients with generalized forms of epilepsy (such as major motor or petit mal seizures) have global (that is, throughout the brain) increases in glucose utilization in the ictal period of seizure activity with subsequent depression in glucose utilization in the postictal period (*19*).

The finding of interictal zones of hypometabolism, particularly in the temporal lobe of partial epilepsy patients, has suggested that this may be specific for a lowered threshold or greater susceptibility of these foci to initiation of seizure activity. These findings have in turn promoted a series of parallel studies in animals (glucose utilization, blood flow, morphology, electrophysiology, ligand

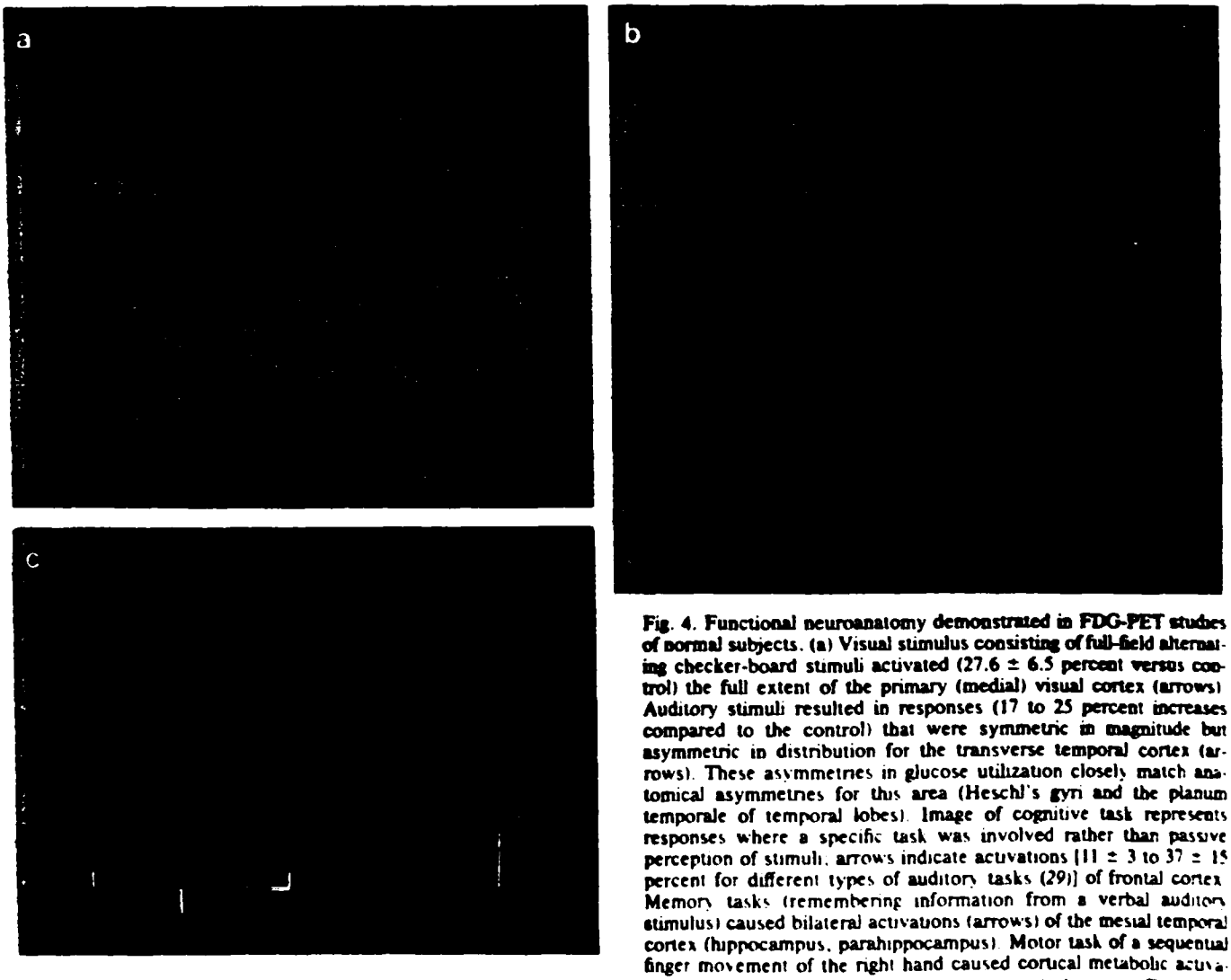


Fig. 4. Functional neuroanatomy demonstrated in FDG-PET studies of normal subjects. (a) Visual stimulus consisting of full-field alternating checker-board stimuli activated (27.6 ± 6.5 percent versus control) the full extent of the primary (medial) visual cortex (arrows). Auditory stimuli resulted in responses (17 to 25 percent increases compared to the control) that were symmetric in magnitude but asymmetric in distribution for the transverse temporal cortex (arrows). These asymmetries in glucose utilization closely match anatomical asymmetries for this area (Heschl's gyri and the planum temporale of temporal lobes). Image of cognitive task represents responses where a specific task was involved rather than passive perception of stimuli, arrows indicate activations [11 ± 3 to 37 ± 15 percent for different types of auditory tasks (*29*)] of frontal cortex. Memory tasks (remembering information from a verbal auditory stimulus) caused bilateral activations (arrows) of the mesial temporal cortex (hippocampus, parahippocampus). Motor task of a sequential finger movement of the right hand caused cortical metabolic activation (18.6 ± 3.9 compared to the control) of the left motor strip (lower arrow) and supplementary motor cortex (vertical arrow). Errors are standard deviations. [Courtesy of M. E. Phelps and J. C. Mazziotta, UCLA School of Medicine] (b) Auditory stimuli produced metabolic responses that varied with the content and in some cases strategy used by the subject to perform the task. In resting states (ears plugged, eyes open) left-right cerebral symmetry (left/right = 1.01 ± 0.03) in glucose utilization is seen. Verbal auditory stimuli predominantly activated and caused metabolic asymmetries (left > right = 5 to 16 percent) of the left hemisphere while nonverbal stimuli (music) predominantly activated the right hemisphere (16 to 27 percent) in right-handed individuals. Simultaneous stimulation with language and music caused bilateral activations of both hemispheres. [From J. C. Mazziotta *et al.* (*29*), courtesy of *Neurology*] (c) Nonverbal auditory stimuli caused activations of the language nondominant (right) hemisphere. The timbre test required subjects to compare complex chords for similarities and differences and consistently produced increases (22 percent versus control) in glucose utilization in the right hemisphere (image at right, arrow). This test did not contain any temporal sequences of notes that can be analytically perceived. The chord pairs differed only in tonal quality. When subjects were asked to identify differences and similarities in sequences of notes (tone sequences) the side of maximal activation correlated with the strategy used by the subject to perform the task. Color scale is in units of micromoles per minute per 100 g, ranging from 2 (dark purple) to 45 (red). See text for description. [Modified from J. C. Mazziotta *et al.* (*29*), courtesy of *Neurology*]

assays (*20*) and in humans (glucose utilization, oxygen metabolism, and blood flow) with PET in ictal, postictal, and interictal states (*17–19, 21*) along with morphological and ligand studies in surgically resected tissue samples. This illustrates the manner in which PET, animal, and laboratory assays can be assembled to study the underlying mechanisms of a human disorder.

In drug-free patients with bipolar (manic-depressive) mood disorders, measurements with FDG and PET have shown changes throughout the brain in glucose utilization during different mood states (Fig. 2, b and c). During the depressive phase of illness, patients demonstrated reductions in glucose utilization of about 25 percent throughout the supratentorial structures of the brain which were significantly ($P < 0.001$) below those of age- and sex-matched normals (*22*).

The same patients, when later studied in hypomanic (state between normal and mania) phases of the illness, showed glucose utilization rates that were not significantly different from those of age- and sex-matched controls. Although the major changes in glucose utilization seen thus far have been global, the largest differences between behavioral states occurred in the frontal and anterior cingulate cortices (*22*). Patients with unipolar mood disorders (depression) in drug-free states were found to have global supratentorial cerebral glucose utilization values that were not significantly different from age- and sex-matched controls. However, these patients did exhibit significantly depressed glucose utilization (15 percent lower than age- and sex-matched normals) bilaterally in the striatum (*22*). This reduced striatal glucose utilization recovered to normal levels in patients who became euthymic (that is, normal mood) spontaneously or by drug therapy (*22*).

Physiological psychology. Many methodologies have been used in the study of normal cerebral function. These approaches have included neuropsychological, electrophysiological, and behavioral observations of human subjects performing various tasks. Studies in animals by means of biochemical, electrophysiological, and autoradiographic techniques have also been used to define the anatomical and functional organization of the brain during defined tasks. In a similar manner, PET provides a means of studying local sensory, motor, memory, and cognitive function in normal subjects in vivo (Fig. 4). Thus, much in the way Penfield and his colleagues (*23*) mapped cerebral function through intraoperative stimulation of the human cerebral cortex, PET can provide physiological and biochemical information about normal cerebral function for all human brain regions in a noninvasive fashion.

Various sensory and motor stimulation tasks (Fig. 4) have been used to define the functional neuroanatomy of the human brain. Visual stimulation studies, with ^{15}O-labeled water to measure cerebral blood flow, or with FDG to measure glucose utilization have revealed some of the normal physiological response characteristics of the human visual cortex (*24–28*). These studies have shown (i) the topography of the human visual cortex relative to the site and size of retinal stimulation (*24*), (ii) that the magnitude of blood flow or glucose utilization of the visual cortex is a function of stimulus complexity (*25, 26*) and rate (*16*), (iii) that functionally 50 percent of the input from each eye goes to each visual cortex (*25, 26*), and (iv) that lesions of the visual system (both within and outside the visual cortex) produce

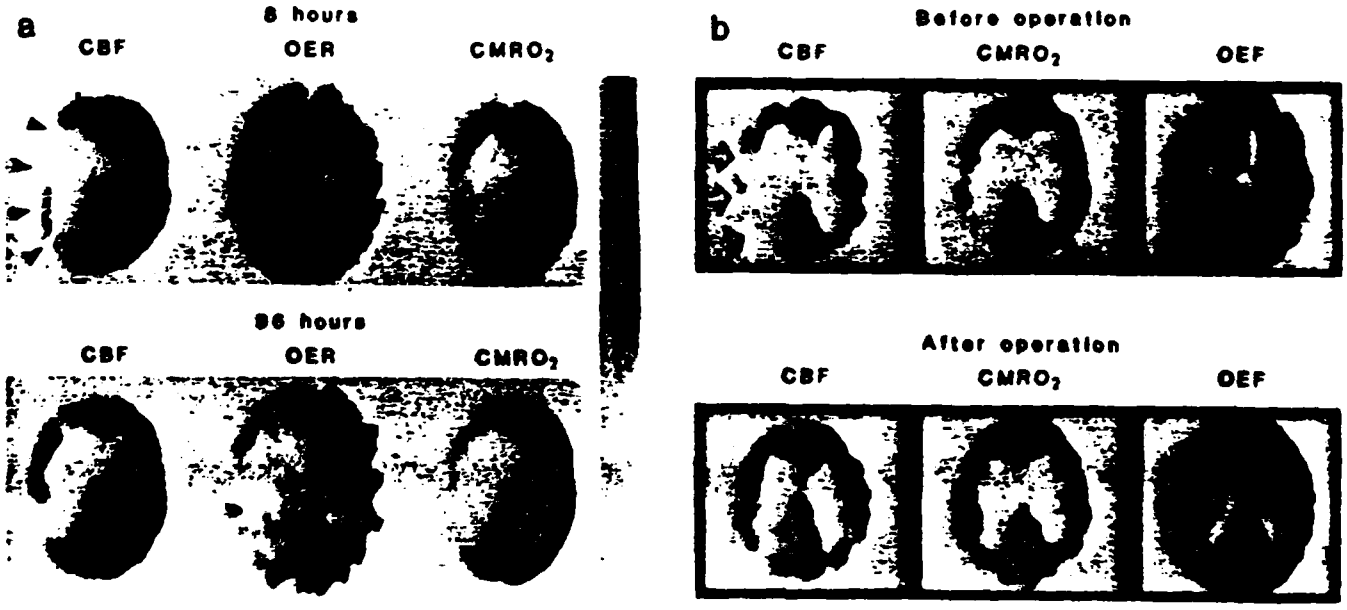


Fig. 5. Pathophysiology of cerebrovascular disease. (a) A patient with left hemisphere cerebral ischemia studied 8 and 96 hours after the onset of symptoms. During both intervals cerebral blood flow (CBF), oxygen extraction ratio (OER), and oxygen metabolism ($CMRO_2$) were determined. The blood flow 8 hours after the stroke was severely depressed in the left middle cerebral artery distribution. Oxygen utilization was also reduced although not as severely as blood flow. The resultant mismatch accounts for increased oxygen extraction (mean value at 14 hours after symptom onset = 0.71 ± 0.12 (R. J. Wise *et al.*, in *34*), normals = 0.49 ± 0.02 (R. S. Frackowiak *et al.*, in *34*) in the viable but jeopardized zone. At 96 hours, blood flow became heterogenous in the area probably because of failure of vascular autoregulatory mechanisms and oxygen utilization has fallen further. Oxygen extraction is now sharply reduced below normal, indicating irreversible brain injury and predicting infarction for this zone. Errors are standard deviations. [From R. S. Frackowiak and R. J. Wise, in (*34*); courtesy of *Neurology Clinics*] (b) Patient with left internal carotid artery occlusion before and after superficial temporal artery-middle cerebral artery bypass surgery. Preoperatively, patient had symptoms referable to the left hemisphere including clumsiness of the right hand and difficulties with language. Blood flow studies with PET demonstrated moderately severe reductions of flow in the cortex of the left parietal lobe with mild reductions in oxygen utilization. The mismatch between flow and metabolism resulted in an increased oxygen extraction fraction (OEF) (arrow in OEF image), an indication that the tissue is viable although in a precarious state. After bypass surgery the patient's symptoms resolved and the blood flow, oxygen extraction, and oxygen metabolism all returned to normal. Both pre- and postoperatively the x-ray CT images of the patient were normal. [From J. C. Baron *et al.* (*36*); courtesy of *Stroke*]

functional abnormalities that can be correlated with the patients' clinical syndromes despite a lack of anatomical alterations detectable by x-ray CT (*26, 28*) (Fig. 2a).

Studies of the human auditory system with FDG indicate a correlation between the distribution of glucose utilization and the content of the stimulus (Fig. 4b) and, in some cases, the strategy used by the subject to solve the task (Fig. 4c) (*29, 30*). While complex patterns of these responses to auditory stimuli were observed, verbal stimuli caused asymmetric increases in glucose utilization in the left hemisphere in right-handed individuals (Fig. 4b). Nonverbal stimuli, such as musical chords, activated primarily right hemisphere areas particularly in the inferior frontal, parietal, and superior temporal regions (Fig. 4b) (*29, 30*). In subjects who listened to sequences of musical notes (*29*) and were asked to determine whether notes in one sequence differed from those in another, the pattern of glucose utilization correlated with the strategy used by the subject (Fig. 4c). Individuals who used specific visual imagery and analytical strategies ("visualizing" frequency histograms or musical scales in their minds for comparing note sequences) had predominantly left hemisphere asymmetries and increases in glucose utilization in the posterior temporal region. Subjects who did not use this strategy to solve the task but rather used mental "resigning" of the notes had activations in the inferior parietal and temporal-occipital regions of the right hemisphere (Fig. 4c) (*29*).

In general, in studies of subjects with PET, when a specific task was involved rather than the mere passive perception of stimuli, frontal cortical zones showed greater glucose utilization (*29, 30*) (Fig. 4a). In addition, subjects who were asked to recall specific aspects of auditory stimuli had activations of the mesial temporal lobe (hippocampus, parahippocampus) that were never seen in situations where auditory perception without memory tasks were required of the subject.

While the study of normal cerebral function through PET is interesting in itself, the above-mentioned tasks can also be useful in seeking a better understanding and improved differential diagnosis of cerebral disorders (*30*). Thus, much in the way a cardiologist imposes physical exercise on patients to induce detectible changes in altered myocardial function, so too the neuroscientist can induce cerebral work by the use of neurobehavioral or pharmacological cerebral stimuli. This approach may reveal subtle or early cerebral dysfunction that is at or near the limit of functional reserve of the brain.

Stimulation tasks can also be used to investigate cerebral reorganization or compensatory responses after sudden damage to the brain or during ongoing structural degeneration of the brain in certain disorders (*30*). The resulting data should provide clues as to how the brain functionally reorganizes or adapts to perform a task when the system identified for the normal performance of that task has been compromised by cerebral pathology.

Cerebral stoichiometry and pathophysiology of disease. The normal brain has a fairly constant stoichiometry among various substrate utilization rates, and changes in these relationships should be sensitive in detecting and providing insights in the mechanisms of early derangements of cerebral systems induced by disease. Such changes may also be critical in determining the type and timing of treatment. By examining a large number of biochemical and physiological processes in normal subjects with PET, one can define the normal relation between these processes and their range of variability. Once known, these relations can be examined in disease states. Serial studies in patients with cerebral disorders, with multiple tracers to determine cerebral stoichiometry as a function of time, can help identify the pathophysiological sequence of events that occurs during the expression of a syndrome. Similarly, one can examine the stoichiometrically changing relations that occur as the disease advances, as the patient recovers from the disease, or as a result of therapeutic interventions. Some examples of the power of this technique can already be found in studies performed with PET (*31, 33–39*) (Figs. 5 and 6).

Huntington's disease is an inherited

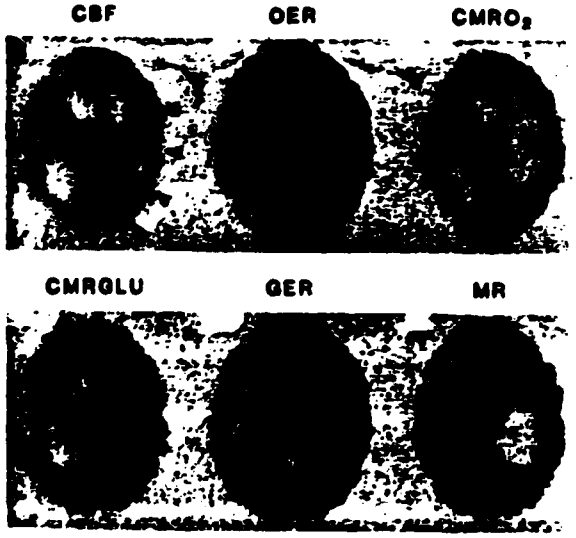


Fig. 6. Abnormal cerebral stoichiometry. (a) Images from a patient with a deep right hemisphere grade IV astrocytoma demonstrating six physiological variables determined with PET. The cerebral blood flow image (CBF) demonstrates a patchy, heterogenous flow in the area of the tumor (large arrow) and diffuse depression in blood flow for the overlying cortex (small arrowheads). The oxygen extraction ratio image (OER) demonstrates decreased oxygen extraction within the tumor itself even though it is well perfused with blood. Oxygen utilization image ($CMRO_2$) demonstrates very low oxygen metabolism within the tumor and diffuse depression of oxygen utilization in the overlying cortex. Mismatch between flow and oxygen utilization within the tumor is evidenced by the low oxygen extraction for that area. The matched reduction in flow and oxygen metabolism in the overlying cortex results in a relatively normal oxygen extraction fraction for that zone. Glucose utilization (CMRGlu) within the tumor is increased relative to surrounding brain. Glucose extraction ratio (GER) is slightly increased in the area of the tumor (*48*). [Courtesy of MRC Cyclotron Unit, Hammersmith Hospital, London, United Kingdom] (b) Studies from a patient with a deep right hemisphere astrocytoma. The ^{68}Ga-labeled EDTA image demonstrates very minimal change in permeability of the blood-brain barrier to this agent. The [^{11}C]glucose image demonstrates a lower than normal utilization rate for glucose within the tumor (arrow). The [^{11}C]methionine image reveals a marked increase in methionine uptake within this tumor. An increased use of amino acids for metabolism and protein synthesis provides an additional stoichiometric variable by which to evaluate the pathophysiology of tumor growth and possible therapeutic interventions. [From M. Bergstrom, in (*38*); courtesy of *Journal of Computer-Assisted Tomography*]

disorder manifested by dementia, abnormal movements and psychiatric symptoms. The offspring of affected individuals have a 50 percent chance of inheriting the abnormal gene and manifesting the disease. The clinical manifestations typically do not appear until the third or fourth decade of life. Thus, a population exists that is at risk for the disease. These individuals may have subclinical expression of symptoms although previous attempts to identify such presymptomatic patients have been, until recently, unreliable.

In PET studies with FDG, all patients with Huntington's disease showed a reduction in glucose utilization in the striatum (up to 70 percent reductions), the portion of the brain with the most profound structural changes (neuronal cell loss) in the advanced stages of this disorder (*31*). In those patients who were just beginning to manifest symptoms of this disease, structural imaging techniques such as x-ray CT were normal (Fig. 3b), but investigations with FDG and PET revealed profound reductions in glucose utilization in the striatum (*31*).

In a group of 15 asymptomatic individuals who were at-risk for Huntington's disease, about half demonstrated mild-to-moderate (up to 40 percent) reductions in glucose utilization for the caudate nucleus (Fig. 3c) (*31*). Three of the individuals with these abnormalities subsequently developed symptoms (*31*). Thus, PET can identify functional lesions in Huntington's disease that precede gross structural changes in the brain (as determined by x-ray CT) and appears to be able to identify abnormalities in patients even before symptoms of the disease are manifest (namely, changes that are still within the compensatory mechanisms of the brain). The relation of these presymptomatic changes in glucose utilization with the presence of the abnormal gene (*32*) remains to be determined.

Serial studying of at-risk individuals with various positron labeled tracers from early ages through the onset of symptoms can provide a detailed description of the regional changes in brain biochemistry as a function of time along with a correlation of these changes with clinical symptoms. Since experimental treatments for such disorders would be most successful when used in presymptomatic or early stage subjects, PET may be useful in identifying those subjects and in providing more objective evidence as to whether such therapeutic interventions are beneficial or harmful.

Patients with cerebrovascular disease (Fig. 5) have been studied with PET to understand better the relationship of cerebral blood flow with oxygen metabolism and extraction (*16, 33–37*). These studies have provided some initial insights into the compensatory mechanisms used by the brain to maintain tissue viability despite decreased substrate availability. As blood flow to the brain is decreased (in the initial minutes and hours after the onset of stroke symptoms), the percentage of arterial oxygen extracted into cerebral tissue increases to a maximal level in order to maintain oxidative metabolism (*34*). Subsequently, the capacity of this and other compensatory mechanisms can be exhausted, and if so, oxygen metabolism will fall. This appears to occur at an oxygen utilization rate of about 0.58 micromoles per minute per gram of tissue (*35*). Along with this change, the percentage of oxygen extracted by cerebral tissue also declines, an indicator of irreversible tissue damage (Fig. 5). The measurement of oxygen extraction and utilization has proved to be a much more reliable predictor of tissue degeneration than blood flow alone since the latter can be decreased, normal, or even increased (reactive hyperemia) at different stages or times in the progressive development of cerebral ischemia and infarction (*16, 33, 34*). These measurements have been combined to select patients for therapeutic intervention and to monitor the effectiveness of the treatment (Fig. 5) (*36, 37*).

Patients with brain tumors have been studied with several positron labeled tracers (Fig. 6) in order to (i) assess their pathological stoichiometry (*38, 39*), (ii) identify differences in these relations for different tumor grades (*39*), and (iii) predict and evaluate the effects and response of a given tumor to a specific radio- or chemotherapeutic modality (*40*). The combined knowledge of energy

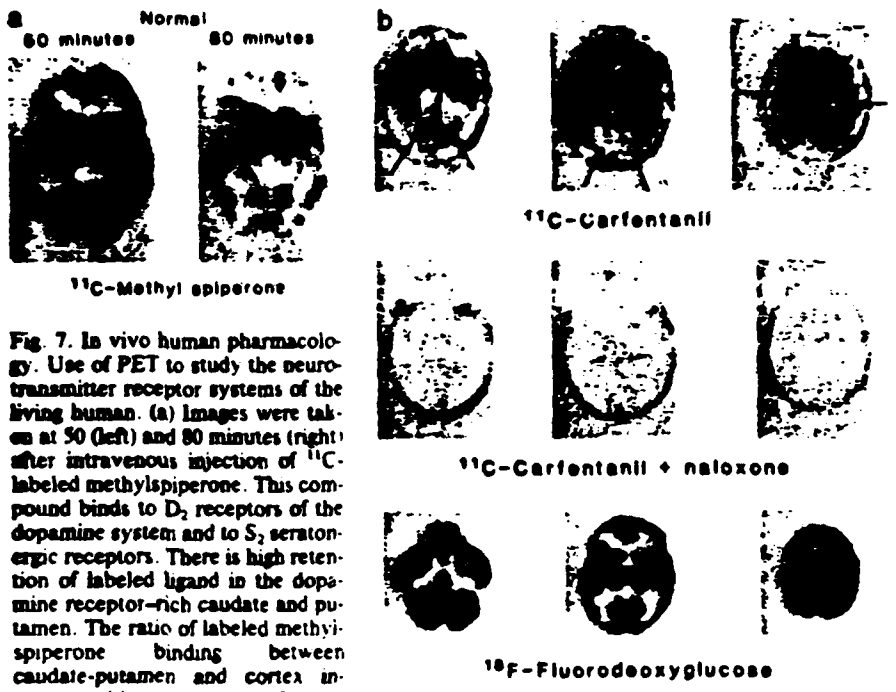


Fig. 7. In vivo human pharmacology. Use of PET to study the neurotransmitter receptor systems of the living human. (a) Images were taken at 50 (left) and 80 minutes (right) after intravenous injection of ^{11}C-labeled methylspiperone. This compound binds to D_2 receptors of the dopamine system and to S_2 serotonergic receptors. There is high retention of labeled ligand in the dopamine receptor–rich caudate and putamen. The ratio of labeled methylspiperone binding between caudate-putamen and cortex increases with time because of retention on the D_2 sites as compared to clearance from nonspecific binding sites and S_2 receptors in the cortex. Ratio of radioactivity in caudate to cerebellum in right-hand image is 4.4. This study also illustrates the high sensitivity of PET to detect concentrations of picomoles per gram or less. [From Wagner *et al.*, in (*42*)] (b) Opiate receptors in human brain. The PET images in the top row were obtained 30 to 60 minutes after intravenous administration of 25 mCi of [^{11}C]carfentanil (80 ng/kg), a mu opiate antagonist. The three images are 7.2 cm, 4 cm, and 0.8 cm above the canthomeatal line (far right to left). Images in the middle row were acquired 30 to 60 minutes after intravenous administration of (1 mg/kg) naloxone (the + isomer), which is an opiate antagonist, and the same dose of [^{11}C]carfentanil used in the first study. In the top row a preferential accumulation of activity is seen in areas rich in opiate receptors such as the thalamus, basal ganglia, and frontal cortex (center and right-hand images) and pituitary gland (left-hand image; inner arrow). Low activity is seen where opiate receptors exist in low concentration, such as the occipital cortex (center image, arrows), the postcentral gyrus (right-hand image; arrows) and the cerebellum (left-hand images; arrows). Images in the middle row demonstrate the low level of nonreceptor binding when labeled carfentanil binding is blocked with naloxone. Approximately 90 percent of specific opiate receptor binding in the thalamus and basal ganglia is displaced. The outer rings in the images result from ^{11}C activity in scalp. Images in the bottom row represent glucose utilization (FDG) for approximately the same levels as the opiate receptor distribution study. [Carfentanil studies from J. J. Frost *et al.*, in (*43*) and glucose utilization images from M. E. Phelps *et al.*, in (*3*)]

metabolism and protein synthesis for assessing cell turnover rates and tumor growth before and after therapy can be evaluated with PET (Fig. 6). For example, radiation therapy is more effective in patients with tumors having high oxygen concentrations (because of the generation of free radicals) than in patients with tumors having low oxygen concentrations (Fig. 6). The determination of the degree of malignancy of cerebral gliomas in vivo with PET measurements of glucose utilization correlates well with histological grading of biopsied or resected samples—that is, an increasing degree of malignancy was associated with increasing rates of glucose utilization as measured with FDG (*39*).

The reliability and appropriateness of animal models of human diseases are frequently questioned. The ability to compare animal models and human conditions through, for example, biochemical assays with the tracer kinetic approach provides the opportunity to scrutinize these animal models and compare differences and similarities with the human disease. Animal models of different aspects of epilepsy, neoplastic diseases, psychiatric disorders, Parkinson's and Huntington's diseases, and others can be examined by quantitative autoradiography and biochemical assays for comparison with PET studies of patients having these disorders. Similarities and differences in stoichiometry, structure-function relationships, behavior-function relationships, and drug responses with PET can help in the choice of appropriate models. The selected animal model can then be studied by histological, biochemical, and electrophysiologic techniques that are either too invasive or too logistically complex to be performed in humans. Hypotheses resulting from studies on mechanisms of the disease in animals can then be tested in humans with PET.

In vivo pharmacology. The neuropharmacology of the human brain can be examined with PET through the use of two different strategies. In the first, pharmacological doses of drugs can be administered to normal subjects and patients with cerebral disorders and their effect on processes such as blood flow and metabolism can be examined to determine the anatomical sites where drug-induced alterations in biochemical processes occur (*41*). Alternatively, the drug itself can be labeled with a positron-emitting isotope (Table 1) and its pharmacokinetic behavior (*42–45*) can be examined directly in vivo under conditions where the drug is present in tracer amounts (that is, with no mass effect) or in concentrations where pharmacological effects are produced (Fig. 7). Either way, biochemical assays can be performed to examine the pharmacological effects of specific agents on behavior, symptoms, and structure-function relationships in the human brain and to identify neurochemical systems involved in specific diseases. The anatomical sites of these effects can be correlated with neurochemical systems associated with such sites in the brain as determined through in vitro studies of human tissue or animal studies of these systems.

The ability to observe in vivo drug pharmacokinetics can be useful in identifying the responsiveness of different patient groups to specific pharmacologically active agents. For example, let us consider a presumed clinically homogeneous population of patients in which a specific drug is effective in alleviating symptoms in 15 percent. Such a therapy would be considered relatively ineffective for that population as a whole. However, for the 15 percent that responded, the therapy is quite effective. Conventional techniques may not be sensitive enough to identify subpopulations with specific pharmacological sensitivities to different agents. PET studies with either of the two strategies outlined above could provide additional diagnostic information that could potentially segregate such subgroups and allow for more specific and pathophysiologically appropriate therapies to be employed.

A number of labeled compounds have been developed for use with PET in the study of dopaminergic (*42*), opiate (*43*), benzodiazepine (*44*), and other systems (*45*) (Table 1). Labeled ligands of high specific activity have been shown with PET to reflect specific receptor systems as determined by their localization in appropriate anatomical sites, competitive blockade, and kinetic differentiation of specific and nonspecific binding in serial measurements (*42–45*) (Fig. 7). The use of appropriate positron-labeled compounds makes it possible to examine the presynaptic site of neurochemical transmission with a labeled precursor of an active agent such as ^{18}F-labeled L-dopa (*42*), or to examine postsynaptic receptor interactions with an agent such

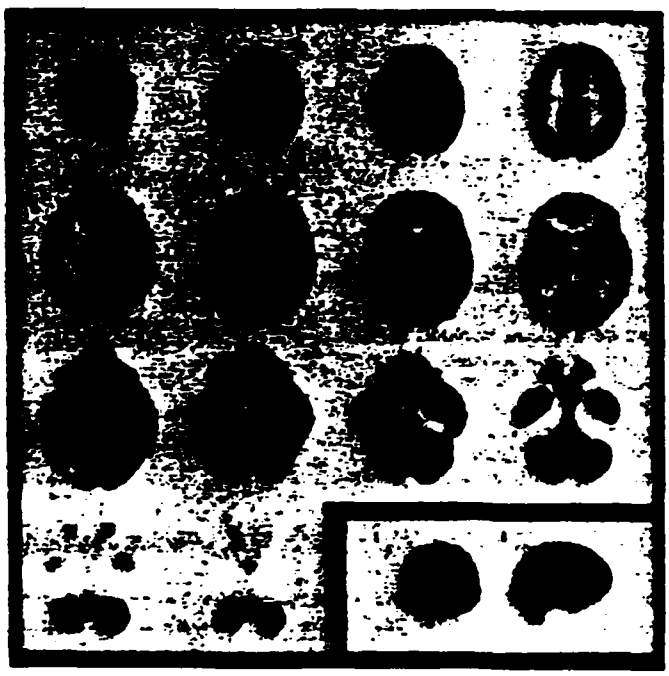


Fig. 8. Complete set of FDG-PET images from a normal subject, demonstrating glucose utilization. These images were obtained with an 8-mm interval between slices with the Neuro-ECAT system (image spatial resolution of 8.4 by 8.4 by 12.5 mm). The two images in the lower right corner are anterior-posterior and lateral two-dimensional views (rectilinear studies) of the same subject obtained prior to the tomographic examination. The PET device provides identification of the position of the tomographic planes and allows superimposition of their location on the rectilinear images. Shades of gray, units of micromoles per minute per milligram, with black representing the highest value. At this spatial resolution, the folds in the cortical ribbon are clearly delineated, as are the subcortical structures (right-most images, second row) including the thalamus, caudate, and the lenticular nuclei (putamen-globus pallidus complex). Both the posterior and the anterior limbs of the internal capsule are visualized. The hippocampus-parahippocampal gyri region can be seen lateral to the brain stem, along the medial portions of the temporal lobes (third row, extreme left). Substructures of the posterior fossa are visible (third row, extreme right) and include the brainstem and the substructures of the cerebellum (the cerebellar cortex, vermis, and dentate nuclei). This study demonstrates the progressive increase in image quality that has occurred as the spatial resolution of PET devices has improved. For example, compare the image quality and detail in Figs. 3, b and c, 4 and 5, obtained with a tomograph having a spatial resolution of 16 by 16 by 18 mm with those in this figure. Tomographs with 2- to 4-mm resolution are now being tested. Improvements in spatial resolution not only increase the anatomical detail, the structure identification capacity, and the quantitative accuracy. [From M. E. Phelps *et al.*, in (*3*); courtesy of *Journal of Cerebral Blood Flow and Metabolism*]

as ^{11}C-labeled N-methylspiperone (42), or ^{18}F-labeled spiperone (42) for the (D_2) dopaminergic sites, or ^{11}C-labeled carfentanil (43) for the (mu) opiate sites (Fig. 7). To transform these types of studies into measurements of processes such as receptor affinity and density, neurotransmitter concentration and turnover, ligand on and off rates, and diffusion rates requires formulation and kinetic and biochemical verification of models, whether for PET or any in vivo estimation (3, 46).

Conclusion

The examples of the use of PET to examine structure-function relationships, physiological psychology, pharmacology, stoichiometry, and pathophysiology of disease illustrate some of the diverse types of investigational strategies that can be utilized with this technology. The further course of development of PET depends not only on advances in the technology itself (Fig. 8) but also on its role in using and applying advances from the basic biomedical sciences to human studies. This is aided by the many types of labeled compounds already available, or that can be synthesized, providing PET with a broad range of biochemical and physiological probes for examining different processes of the brain. Required advances in the number and validity of tracer kinetic models continue not only in the field of PET but in all those fields where radioactive labeled tracers are used. The importance of an analogous scientific method for studies in man to those in a laboratory setting, underscores the significance and potentials of PET.

References and Notes

1. M. E. Phelps, E. J. Hoffman, N. A. Mullani, M. M. Ter-Pogossian, *J. Nucl. Med.* 16, 210 (1975).
2. M. E. Raichle, *Brain Res. Rev.* 1, 47 (1979); M. M. Ter-Pogossian, M. E. Raichle, B. E. Sobel, *Sci. Am.* 243, 170 (1980).
3. M. E. Phelps, J. C. Mazziotta, S. C. Huang, *J. Cereb. Blood Flow Metab.* 2, 113 (1981); M. Walker, *Ann. Neurol.* 15 (suppl.), S1 (1984); W. D. Heiss and M. E. Phelps, *Positron Emission Tomography of the Brain* (Springer-Verlag, New York, 1983).
4. J. S. Fowler and A. P. Wolf, in *Nuclear Sciences Series*, NAS-NS-3201 (National Academy of Sciences–National Research Council, Washington, D.C., 1983); M. J. Welch and J. P. Tewson, in *Radiopharmaceuticals II*, J. A. Sorenson, Ed. (Society of Nuclear Medicine, New York, 1979), pp. 201.
5. L. Sokoloff *et al.*, *J. Neurochem.* 28, 897 (1977); L. Sokoloff, *J. Cereb. Blood Flow Metab.* 1, [illegible] (1981); *Ann. Neurol.* 15 (suppl.), S1 (1984).
6. E. M. Bessel, A. B. Foster, J. H. Westwood, *Biochem. J.* 128, 199 (1972); B. M. Gallagher *et al.*, *J. Nucl. Med.* 19, 1154 (1978); E. E. Machado de Domenech and A. Sols, *FEBS Lett.* 119, 174 (1980).
7. M. E. Phelps *et al.*, *Ann. Neurol.* 6, 371 (1979); S. C. Huang *et al.*, *Am. J. Physiol.* 238, E69 (1980).
8. M. Reivich *et al.*, *Circ. Res.* 44, 127 (1979).
9. R. Hawkins, M. E. Phelps, S. C. Huang, D. E. Kuhl, *J. Cereb. Blood Flow Metab.* 1, 37 (1981).
10. F. Schuier *et al.*, *ibid.* 1 (suppl. 1), S63 (1981).
10a. B. K. Siesjo, personal communication.
11. M. E. Phelps, S. C. Huang, J. C. Mazziotta, R. A. Hawkins, *J. Cereb. Blood Flow Metab.* 3 (suppl. 1), S131 (1983); A. Gjedde and N. H. Diemer, *ibid.*, p. 303; W. M. Pardridge *et al.*, *ibid.* 2, 197 (1982).
12. S. S. Kety, *Pharmacol. Rev.* 3, 1 (1951).
13. B. K. Siesjo, *Brain Energy Metabolism* (Wiley, New York, 1978), p. 101.
14. R. A. Hawkins, W. K. Hass, J. Ransohoff, *Stroke* 10, 690 (1979).
15. M. E. Raichle *et al.*, *Am. J. Physiol.* 228, 1936 (1975).
16. R. H. Ackerman *et al.*, *Arch. Neurol.* 38, 537 (1981).
17. J. Engel, Jr., D. E. Kuhl, M. E. Phelps, J. C. Mazziotta, *Ann. Neurol.* 12, 510 (1982); D. E. Kuhl, J. Engel, Jr., M. E. Phelps, C. Selin, *ibid.* 8, 348 (1980); J. C. Mazziotta and J. Engel, Jr., *Epilepsia* 25 (suppl. 2), S86 (1984); W. H. Theodore *et al.*, *Ann. Neurol.* 14, 419 (1983).
18. J. Engel, Jr., D. E. Kuhl, M. E. Phelps, R. Rausch, M. Nuwer, *Neurology* 33, 400 (1983).
19. J. Engel, Jr., D. E. Kuhl, M. E. Phelps, *Science* 218, 64 (1982).
20. R. F. Ackermann, D. M. Finch, T. L. Babb, J. Engel, Jr., *J. Neurosci.* 4, 251 (1984); Y. Ben Ari, E. Tremblay, D. Riche, E. Ghilini, R. Naquet, *Neuroscience* 6, 1361 (1981); W. F. Caveness *et al.*, *Ann. Neurol.* 7, 213 (1980); R. C. Collins, *Brain Res.* 150, 487 (1978); R. E. Watson *et al.*, *Brain Res. Rev.* 5, 133 (1983).
21. S. Bernardi, M. R. Trimble, R. S. Frackowiak, R. J. Wise, T. Jones, *J. Neurol. Neurosurg. Psychiatr.* 46, 473 (1983).
22. M. E. Phelps, J. C. Mazziotta, L. Baxter, R. Gerner, *Ann. Neurol.* 15 (suppl.), S192 (1984); L. R. Baxter *et al.*, *Arch. Gen. Psychiatr.* 42, 1985.
23. W. Penfield and T. Rasmussen, *The Cerebral Cortex of Man* (Macmillan, New York, 1950).
24. J. Greenberg *et al.*, *Science* 212, 678 (1981); M. Kushner *et al.*, *Ann. Neurol.* 12, 89 (1982); J. C. Mazziotta, M. E. Phelps, E. Halgren, *Human Neurobiol.* 2, 11 (1983); J. C. Mazziotta *et al.*, *J. Cereb. Blood Flow Metab.* 3 (suppl. 1), S246 (1983); M. Reivich *et al.*, *Acta Neurol. Scand.* 60 (suppl. 72), 198 (1979); E. L. Schwartz, D. R. Christman, A. P. Wolf, *Brain Res.* 294, 225 (1984).
25. M. E. Phelps, D. E. Kuhl, J. C. Mazziotta, *Science* 211, 1445 (1981).
26. M. E. Phelps *et al.*, *Neurology* 31, 517 (1981).
27. P. T. Fox and M. E. Raichle, *J. Neurophysiology* 51, 1109 (1984).
28. M. Reivich *et al.*, *J. Cereb. Blood Flow Metab.* 1 (suppl. 1), S471 (1981).
29. J. C. Mazziotta, M. E. Phelps, R. E. Carson, D. E. Kuhl, *Neurology* 32, 921 (1982).
30. J. C. Mazziotta and M. E. Phelps, *Ann. Neurol.* 15 (suppl.), S50 (1984).
31. D. E. Kuhl *et al.*, *ibid.* 12, 425 (1982); D. E. Kuhl *et al.*, *Brain Imaging and Brain Function* (Raven, New York, in press).
32. J. Gusella *et al.*, *Nature (London)* 306, 234 (1983).
33. R. H. Ackerman *et al.*, *J. Cereb. Blood Flow Metab.* 1 (suppl. 1), S502 (1981); J. C. Baron *et al.*, *Trans. Amer. Neurol. Ass.* 105, 459 (1980); J. C. Baron *et al.*, *Acta Neurol. Scand.* 60 (suppl. 72), 324 (1979); G. L. Lenzi *et al.*, *Eur. Neurol.* 20, 285 (1981); W. R. Martin and M. E. Raichle, *Ann. Neurol.* 14, 168 (1983).
34. J. C. Baron, M. G. Bousser, D. Comar, F. Soussaline, P. Castaigne, *Eur. Neurol.* 20, [illegible] (1981); R. S. J. Frackowiak and R. J. Wise, *Neurol. Res.* 1, 189 (1983); R. J. Wise, S. Bernardi, R. S. J. Frackowiak, T. Jones, *Brain* 106, 197 (1983); R. J. Wise, *Ann. Neurol.* 14, 63[illegible] (1983); R. S. J. Frackowiak *et al.*, *J. Comput. Assist. Tomogr.* 4, [illegible] (1980).
35. C. L. Lenzi, R. S. J. Frackowiak, T. Jones, *J. Cereb. Blood Flow Metab.* 1 (suppl. 1), S[illegible] (1981).
36. J. C. Baron *et al.*, *Stroke* 12, 454 (1981).
37. J. M. Gibbs *et al.*, *J. Cereb. Blood Flow Metab.* 3 (suppl. 1), S590 (1983); W. R. Martin *et al.*, *Neurology* 32 (suppl. 2), A[illegible] (1982); W. J. Powers, W. R. Martin, P. Herscovitch, M. E. Raichle, R. Grubb, *J. Nucl. Med.* 24, P10[illegible] (1983); Y. L. Yamamoto, S. Little, C. Thompson, E. Meyers, J. V. Feinde, *Acta Neurol. Scand.* 60 (suppl. 72), 522 (1979).
38. M. Bergstrom, *J. Comput. Assist. Tomogr.* 7, 1062 (1983); M. Ito *et al.*, *Neuroradiology* 23, 63 (1982); C. G. Rhodes *et al.*, *Ann. Neurol.* 14, 614 (1983).
39. G. DiChiro *et al.*, *J. Cereb. Blood Flow Metab.* 1 (suppl. 1), S11 (1981); G. DiChiro *et al.*, *Neurology* 32, 1323 (1982); G. DiChiro *et al.*, in *Positron Emission Tomography of the Brain*, W. D. Heiss and M. E. Phelps, Eds. (Springer-Verlag, New York, 1983), p. 182; G. DiChiro *et al.*, *J. Comput. Assist. Tomogr.* 7, 93[illegible] (1983); P. L. Kornblith *et al.*, in *Progress in Experimental Brain Tumor Research*, M. L. Rosenblum and C. B. Wilson, Eds. (Karger, Basel, in press).
40. N. J. Patronas *et al.*, *Radiology* 144, 885 (1982).
41. M. E. Phelps, J. C. Mazziotta, R. Gerner, L. Baxter, D. E. Kuhl, *J. Cereb. Blood Flow Metab.* 3 (suppl. 1), S[illegible] (1983); J. McCulloch, in *Handbook of Psychopharmacology*, L. L. Iverson, S. D. Iverson, S. H. Snyder, Eds. (Plenum, New York, 1982), p. 321.
42. E. S. Garnett, G. Firnau, C. Nahmias, *Nature (London)* 305, 137 (1983); K. L. Leenders *et al.*, *Lancet* 1984-II, 110 (1984); J. C. Baron *et al.*, in *Positron Emission Tomography of the Brain*, W. D. Heiss and M. E. Phelps, Eds. (Springer-Verlag, New York, 1983), p. 212; H. N. Wagner, Jr., *et al.*, *Science* 221, 1264 (1983).
43. M. Maziere *et al.*, *J. Labeled Compd. Radiopharmacol.* 18, 15 (1981); J. J. Frost *et al.*, *J. Nucl. Med.* 25, P73 (1984).
44. D. Comar, M. Maziere, J. M. Gado, G. Berger, Fg. Soussaline, *Nature (London)* 280, 329 (1979).
45. J. C. Baron, *Neurology* 33, 580 (1983).
46. M. A. Mintun, M. E. Raichle, M. R. Kilbourne, G. F. Wooten, M. J. Welch, *Ann. Neurol.* 15, 217 (1984).
47. D. E. Kuhl, *Radiology* 150, 625 (1984).
48. In normal brain about 5.6 mol of oxygen are used per mole of glucose. PET measurements of oxygen and glucose utilization allow calculations of this ratio in patient studies. The tumor in Fig. 6a (arrow) has a significant reduction in this ratio (value of about 1.9) indicating its dependence on anaerobic glycolysis despite adequate oxygen availability through its elevated blood flow. Although there is depressed function of the cortex on the right in this patient, the molar ratio of oxygen to glucose is normal (about 5.4), an indication that this is unlikely to be a result of ischemia from pressure effects or direct tumor infiltration of the cortex. Rather, this change most likely represents a functional drop in neuronal activity due to disruption of fiber tracts (as was seen in Fig. 1). This type of information (blood flow, anaerobic or aerobic status) is an important factor in planning radio- or chemotherapy of neoplastic lesions.

49. We thank the many investigators who contributed their time and data to this article. Ron Sumida for technical assistance, and Lee Griswold for preparing illustrations. Supported in part by DOE contract AM03-76-SF00012, NIH grants RO1-6M-24838, PO1-NS-15654, NIMH grant RO1-MH-37916-02, grants from the Hereditary Disease and Wills Foundations, donations from the Jennifer Jones Simon Foundation, and NINCDS (J. C. M.) Teacher-Investigator Award 1K07-00588-04-NSPA.

5. Structure/Function Relations

J. Cereb. Blood Flow and Metab.
5 (suppl): In press.

PROBLEMS AND STRATEGIES IN FUNCTIONAL CEREBRAL IMAGE ACQUISITION AND ANALYSIS. John C. Mazziotta, Stephen H. Koslow*, UCLA School of Medicine, Los Angeles, CA, *National Institute of Mental Health, Rockville, MD.

Human tracer kinetic, tomographic imaging that provides three-dimensional cerebral distribution maps of physiologic processes present difficult problems in data acquisition, analysis and presentation. The ultimate goal of these functional cerebral imaging studies is to provide quantitative data on functional states or the localization of exogenous compounds in specific brain nuclei. The functional imaging (FI) field has grown rapidly in the last decade in terms of the number of investigators, the power of its resolution and the scope of its applications. Both the scientific community and the public expect incremental increases in our knowledge of normal and abnormal brain function because of the unique capabilities now available to study the brain in its intact functional state. Our ability to take full advantage of the power of functional imaging and to maintain credibility requires standardized, reproducible and accurate methods of obtaining, analyzing and reporting the resultant data as well as recognition of the limits of these approaches.

In its broadest sense these problems include all aspects of FI acquisition, analysis and presentation. A survey of these issues results in a formidable list. Some of these items, grouped in four separate disciplines are presented in Table 1. For each discipline a set of problems and solutions have been examined at an initial series of workshops sponsored by the National Institutes of Mental Health in the U.S. Representatives of the fields of physics, biomathematics, neuroanatomy and computer science have and will develop position papers describing these results for distribution, constructive criticism and further input from the FI community.

Anatomical issues remain the least defined. From its beginnings, the FI field has related functional images to discrete anatomical brain nuclei. As spatial resolution improves, this is done with greater frequency and confidence. In reality, however, these images represent functional data which is not equivalent to structure and caution needs to be taken in the use of such approaches. Functional images (e.g., PET, SPECT) could be individually paired with structural images (e.g., MRI) but this adds cost, logistical complexity, greater patient compliance and another layer of data acquisition and analysis with their attendant errors. While paired data sets may be required for certain FI applications, some initial optimal criteria for FI studies have been developed which ideally avoid these complexities (Table 2).

A survey of most existing FI analysis approaches has been undertaken and includes: manual template matching, elastic (semiautomatic) template matching from X-ray CT and MRI, stereotactic systems, hybrid combinations of these methods and less anatomically based statistical approaches.

The most promising technique, at present, is based on the use of a stereotactic coordinate system. Such a method can be expanded to include idealized atlas templates that can be three-dimensionally deformed to accomodate differences in individual brain anatomy found in healthy and disease states. It is most probable that such an approach will work in individuals with the normal range of anatomical variabilty. Disease states

would have to be examined individually to determine whether ancillary anatomical data is required.

Test data sets of human brain structures will be required to test this, and potentially other, analysis approaches. Present stereotactic anatomical atlases are derived from single, post-mortem brains, removed from the skull and have only been validated for structures near their coordinate system origin (e.g., thalamus, basal ganglia). It is suggested that a large number of high resolution MRI studies be assembled to develop an anatomical stereotactic in vivo validation atlas. The criteria for collecting this data set are under consideration (e.g., normal vs. representative population, image acquisition variables, race, gender, handedness, sample size, etc). Once available, proposed FI analysis systems could be tested, their accuracy and reproducibility determined and their range of errors established for each anatomical region.

The proposed plan is to develop guidelines for the present and ongoing acquisition of FI data that is compatible with promising FI analysis techniques. Second, pilot data will be examined to understand better the criteria needed to develop the MRI atlas. Third, large data sets will be acquired, the atlas established and the three-dimensional site and variation of brain structures determined. Fourth, the performance of FI analysis schemes will be tested using this information.

While standardization is necessary for communicationn, the goal of such efforts should, in no way, result in constraints on the creativity of the field. Continued and ongoing input from the FI community will be necessary to sustain this effort and refine methods in parallel with our increasing knowledge and advancing technology. The capacity to objectively compare and exchange FI data collected by the resultant methods should justify the efforts and interest of the FI community in these important issues.

A. Physics

1. Instrument quality control (calibration, timing, etc)
2. Spatial resolution
3. Temporal resolution
4. Attenuation correction
5. Accidental events
6. Scatter
7. Deadtime

B. Biomathematics

1. Model behavior for various tracers
2. Parameter estimation
3. Image reconstruction
4. Statistical limitations in data acquisition
5. Data normalization
6. Statistical analysis of data from analyzed images (e.g., left/right, local/global).

C. Neuroanatomy

1. Angle and level of scan acquisition
2. Positioning/Repositioning
3. Landmark identification
4. Variations: gross, microscopic, biochemical, head position (gravitational efforts).
5. Alterations by disease processes

D. Computer Science

1. Data handling
2. Matching of anatomical and functional image sets
3. Computational methods
4. Data presentation (e.g., graphics)
5. Data exchange (within and among centers).

Table 1: Representative disciplines and issues in functional image acquisition and analysis

1. Reproducible
2. Accurate
3. Independent of tracer employed
4. Independent of instrument spatial resolution
5. When possible, independent of ancillary imaging techniques
6. Minimizes subjectivity and investigator bias
7. Fixed assumptions about normal anatomy not required
8. Acceptable to subjects' level of tolerance (head holders, etc).
9. Performs well in serial studies of the same patient and individual study of separate patients in a population
10. Capable of evolving toward greater accuracy as information and instruments improve
11. Reasonable in cost
12. Equally applicable in both clinical and research settings
13. Time efficient for both data acquisition and analysis

Table 2: Proposed criteria for the optimal solution to problems of functional image analysis and acquisition. Practical, logistical and pathologic considerations may limit the degree or number of criteria that can be fulfilled in a given application. Current technological limitations prevent full realization of item 5.

5.2.1

NPRDC TR 84-23

FEBRUARY 1984

BIOTECHNOLOGY PREDICTORS OF PHYSICAL SECURITY PERSONNEL PERFORMANCE: CEREBRAL POTENTIAL MEASURES RELATED TO STRESS

Donald B. Malkoff

APPROVED FOR PUBLIC RELEASE,
DISTRIBUTION UNLIMITED

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

5.2.2

NPRDC TR 84-23

February 1984

BIOTECHNOLOGY PREDICTORS OF PHYSICAL SECURITY PERSONNEL PERFORMANCE: CEREBRAL POTENTIAL MEASURES RELATED TO STRESS

Donald B. Malkoff

Reviewed by
Richard C. Sorenson

Released by
J. W. Renard
Captain, U.S. Navy
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

5.2.3

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPRDC TR 84-23

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): BIOTECHNOLOGY PREDICTORS OF PHYSICAL SECURITY PERSONNEL PERFORMANCE: CEREBRAL POTENTIAL MEASURES RELATED TO STRESS

5. TYPE OF REPORT & PERIOD COVERED: Special Report

6. PERFORMING ORG. REPORT NUMBER: 41-83-04

7. AUTHOR(s): Donald B. Malkoff

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62715H, MPR-83-507

11. CONTROLLING OFFICE NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152

12. REPORT DATE: February 1984

13. NUMBER OF PAGES: 29

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):

Stress
Performance measurements
Evoked fields
Event-related fields
Event-related potentials

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research literature related to stress, personality measurements, electrical and magnetic measurements of brain activity, and stress task-protocols was reviewed to determine whether measurements of brain activity can be used to predict job performance under conditions of stress. Results indicated that brain activity measurements show great promise (1) for predicting general response-tendencies of individuals when subjected to stress and (2) as an investigative method for learning more about brain function. Recommendations were made for a research protocol for ascertaining whether measurements of brain activity can be used to predict job performance under stress.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FOREWORD

This effort was conducted within Defense Nuclear Agency (DNA) program element 62715H (Nuclear Weapons Physical Security Research and Development). The objective of the project is to determine the feasibility of using biotechnology procedures such as bioelectric and biomagnetic recordings of brain activity to improve predictions of physical security personnel reliability and performance effectiveness.

This report is the third in a series resulting from this project. Previous reports provided an annotated bibliography (NPRDC TN 83-9) of the stress literature related to performance and reviewed selected experimental procedures for assessing performance under stress (NPRDC SR 84-9). This report concerns predicting responses to stress using bioelectrical and biomagnetic measurements of brain activity. The problems, advantages, and future implications of this research approach are examined, and recommendations are made regarding implementation. The results should be of interest to the research community concerned with brain functions, human behavior, and performance prediction.

Previous research conducted by the Navy Personnel Research and Development Center on brain event-related potentials is summarized in NPRDC TR 84-3.

J. W. RENARD
Captain, U.S. Navy
Commanding Officer

JAMES W. TWEEDDALE
Technical Director

SUMMARY

Problem

Improved capabilities are desired for measuring and predicting the dependability and performance effectiveness of Navy personnel assigned to security force duties within the nuclear weapons safeguards area, particularly under conditions of stress. Information is needed as to the feasibility of using electrical and magnetic measurements of brain activity (evoked and event-related potentials) for this purpose.

Purpose

The purpose of this effort was to determine whether measurements of brain activity can be used to predict job performance under conditions of stress.

Approach

1. The research literature related to stress, personality measurements, electrical and magnetic measurements of brain activity, and stress task-protocols was reviewed.

2. A research protocol was devised that should be useful in ascertaining whether or not measurements of brain activity can be used as predictors of performance under conditions of stress.

Results and Conclusions

Use of brain activity measurements shows great promise as (1) a tool for predicting general response-tendencies of individuals when subjected to stress and (2) an investigative method for learning more about brain function, particularly as it applies to emotions, human behavior, and individual differences.

Recommendations

Recommendations were made for a research protocol for ascertaining whether measurements of brain activity can be used to predict job performance under stress.

CONTENTS

INTRODUCTION

Problem

The Navy has been interested in developing the ability to predict more reliably the performance of those individuals whose jobs might subject them to acute or chronic "stress." Security guards and aviators are examples of personnel whose performance under stressful conditions can be of crucial importance. Information is needed to determine whether or not electrical and magnetic measurements of brain activity can be used to differentiate individuals who will perform "well" under stress as opposed to those who will perform "poorly," and, if so, to determine how best to conduct such measurements.

Purpose

The purpose of this effort was to determine whether measurements of brain activity can be used to predict performance under conditions of stress.

Background

Previous work at NAVPERSRANDCEN has been directed toward predicting on-the-job performance of individuals by analyzing measurements of their electrical and magnetic brain activity (Lewis, 1983a, 1983b). The bioelectrical measurements have included recordings from subjects of both the spontaneous electroencephalogram (EEG) and the evoked potentials (EPs), in some cases referred to as event-related potentials (ERPs). ERPs differ from EPs in that they result, in part, from internal stimuli, such as cognitive tasks and emotions, rather than from external stimuli alone. The corresponding biomagnetic measurements are the magnetoencephalogram (MEG) and the evoked fields (EFs) or event-related fields (ERFs).

Both EPs and EFs are obtained by presenting to the subject a train of brief stimuli; for example, flashing lights or auditory clicks. The electrical and magnetic activity that are thereby generated within the brain are measured over various areas of the scalp following each stimulus. The activity is averaged together in such a way that random background electrical and magnetic activity, or "noise," is eliminated. The recording of the brain activity "signal" directly resulting from the stimulus is thereby enhanced. The amplitude and latency of this activity may then be analyzed statistically or displayed topographically so as to detect differences between individuals, between separate brain areas, or in the same brain area of an individual at different periods of time.

Ordinarily, the patterns evoked in a specific individual by a stimulus presented under controlled conditions are relatively stable (Hillyard, Picton & Regan, 1978). Measurements of EPs from different subjects, however, even when obtained over the same area of brain and under the same controlled conditions, show significant differences in amplitude and latency patterns (Shagass, 1972b). Because of the considerable evidence in the literature suggesting that specific brain functions tended to be localized to either the left or right brain hemispheres (Gordon, Silverberg-Shalev, & Czernilas, 1982; Kinsbourne, 1978), the Navy Personnel Research and Development Center conducted a number of projects to determine whether ERPs could be used to estimate a subject's potential for mastering certain kinds of skills and, hence, predict his job performance (see Lewis, 1983b).

EPs and ERPs, although relatively stable, do show significant changes as a result of alterations in the internal environment (McCallum, 1979), such as occurs with cognition. Moreover, the changes tend to follow a stereotyped pattern dependent upon the particular environmental change (Begleiter & Porjesz, 1975). Changes such as these can be detected by sequential measurements, consisting of baseline (static) ERPs or EPs, followed by another measurement reflecting the change in a single environmental factor. Such dynamic studies offer exciting research possibilities for the following reason: A single, standardized stimulus such as a flashing light can result in different EP patterns in two individuals. It can be shown, at least empirically, that the patterns have predictive value regarding some aspects of brain function that are involved in job performance. It may be inferred that the different patterns reflect actual differences in the neurophysiological handling of the stimulus and that this, in turn, is correlated with the subject's unique pattern of behavior or skills, at least as it applies to the job in question. If there is some validity to this inference, then the same paradigm could conceivably be applied successfully to other more complex (and even adverse) stimuli in the cognitive (thought processes) and affective (emotional processes) realm, in order to gauge their possible impact upon the subject's future job performance under those same stimulus conditions. There is already ample evidence that cognitive processes profoundly affect EPs (Donchin, Ritter, & McCallum, 1978).

APPROACH

1. The research literature related to stress, personality measurements, electrical and magnetic measurements of brain activity, and stress task-protocols was reviewed.

2. A research protocol was devised that should be useful in ascertaining whether or not measurements of brain activity can be used as predictors of performance under conditions of stress.

RESULTS

Stress and Job Performance

Individual Differences

One individual might behave quite differently than another when both are subjected to the same kind of stressful situation (Hokanson, 1969). One individual might "rise to the occasion" and manifest the ability to cope with and resolve the stress-producing problem, another might perform inadequately or indecisively, while still another might show marked deterioration during stress and literally "fall apart." Whatever the nature of the response, "good" or "bad," effective or counterproductive, it is generally consistent for a given individual, particularly over a short period of time under controlled conditions. Even the terminology used by researchers reflects this observation. In conditions of chronic stress, researchers tend to classify individuals according to their most vulnerable body organ-system, whose dysfunction most severely reflects the effects of the stressor. For example, patients are referred to as "ulcer-prone," "coronary-prone," or "neurodermatitis prone" (Ursin, 1978). Likewise, in acute stress, they tend to classify individuals according to the most prominent emotion they manifest. For example, patients are referred to as "anxiety-prone," "prone to withdrawal," or "having great emotional strength or stability" (Grings & Dawson, 1978). These classification tendencies imply that a

significant degree of consistency exists in a person's repertoire of behavioral responses to stress.

On the other hand, it is also commonly accepted that, over a period of time, a given individual's response to the same stressor can vary significantly in terms of both form and intensity. The degree of intra-individual variability is usually attributed to "conditioning factors," such as the individual's age, genetic makeup, gender, diet, drug intake, pre-existing diseases, and physical and social surroundings; in short, the sum total of his internal and external environment (Selye, 1974). When the response changes, the individual is said to have "adapted," "changed coping mechanisms," "responded to therapy," or been subjected to new or changing stressors (McGrath, 1970b).

"Good" and "Bad" Responses

In view of the complexity of the underlying "conditioning factors" that can mold an individual's final response to a stressor, it is not surprising that there are so many different ways in which individuals in a group vary in their behavior in what appears to be an identical stressful situation. Those resultant responses are often subjected to judgments as to whether they are "good" or "bad" in quality and intensity. Obviously, such judgments are themselves quite subjective--a good response might be good for society or for an employer but bad for the individual, and vice versa (McGrath, 1970a). The soldier who sacrifices his life for his country is said to have served his country well and to have performed (responded) in a good and honorable manner. From the point of view of prolonging his life, however, the response was certainly far from optimum. On the other hand, if detailed information were available about the soldier's psychological makeup and social background, it might well be that his response, even from his personal point of view, was a "good" one. For example, if he felt that such a sacrifice of life was the only response that could lead to the preservation of countless other lives and of ideals that were of far greater importance to him than his own survival, then his dying might well be the optimal response for him personally. Regarding responses to stressors, judgments of good or bad are purely relative, and they are often quite difficult to make.

To predict on-the-job performance of personnel assigned to critical positions, such as security guards, value judgments must not only be made, but must be made before a life-threatening situation occurs.

Predicting Future Job Performance Under Stress

From a practical standpoint, one cannot subject job applicants to real life-threatening situations to assess their qualifications. Paper-and-pencil aptitude tests have been criticized as being ineffective in predicting on-the-job performance (Ghiselli, 1966). However, since the response of personnel in critical positions can be of corresponding critical importance, the possible application of any technique that shows promise in predicting future behavior under stress must be explored. As noted previously, one such technique is ERP/EP measurements, provided that there is an acceptance of the premise that stress, and behavioral responses to stress, are at least in part mediated by central nervous system pathways (the brain) and are thereby accessible to ERP/EP probing.

Stress, Behavior, and the Brain

One has little choice but to acknowledge the role of the brain in mediating behavior of any kind. After all, motor activity, sensory activity, cognition, autonomic activity, circadian rhythms, emotionality, hormonal activity, and the coordination among them all

are either dominantly controlled by or, at the least, inextricably intertwined with the activity of the central nervous system. One of the characteristics of Hans Selye's lifelong work and popularized writings on stress (Selye, 1975) was his constant emphasis upon the adrenal and pituitary hormones and the autonomic nervous system, as opposed to the brain, an emphasis not shared by others (Mason, 1975). Unfortunately, this approach deemphasized the role of the brain and may have been instrumental in retarding, until recently, direct and vigorous research efforts into the central nervous system handling of stress and stress-related behavior.

The Lack of a Definition of Stress

Selye (1973, p. 2) defined stress as "a nonspecific response of the body to any demand." The demand can be pleasant or unpleasant. He further defined the nonspecific response when he described the "general adaptation syndrome," which includes symptoms such as tachycardia, hypothermia, hypotonia, and hypertension. In Selye's view, stress is the result of a "local reaction to a local change," resulting in activation of an unidentified "first mediator" that causes the general adaptation syndrome. The brain is not involved, since "denervated rats still show the classic syndrome when put under stress" and "stress occurs under deep anaesthesia or after deafferentation of the hypothalmus in mammals, as well as in lower forms of life that have no nervous system" (Selye, 1973, p. 6, 9). Claims of that sort by Selye, unconvincing at best, discouraged clinical and basic science research efforts toward elucidating the role of the central nervous system in stress. It was the flag under which an army of publications deluged the stress literature for several decades, looking always and only for the peripheral manifestations of stress such as changes in heart rate, blood pressure, galvanic skin resistance, pupillary size, and many other inconsistent and often irrelevant signs (Trumbull & Appley, 1967; McGrath, 1970c). These came to be regarded as absolute tests for the presence or absence of stress.

Another very typical definition of stress is "the behavioral and physiological response to actual or impending aversive stimuli" (Anisman & Zacharko, 1982, p. 89).

As one can readily see, these two definitions may be contradictory ("pleasant" demands are not adversive); further, when one seeks to obtain further elaboration of key words (e.g., "stimuli," "response," or "demand"), there is none that allows a discriminating consistent, functional definition. Most authors agree that there simply is no adequate single definition of stress (Pepitone, 1967; Hamburg & Elliot, 1981; Anisman & Zacharko, 1982). Indeed, there are those who feel that stress is synonomous with arousal and therefore does not even constitute a separate phenomenon (Mason, 1975). Murison and Ursin (1982, p. 115) offered the following definition: "The simplest operational definition of stress, therefore, is that it is the process which produces a change in your own favorite physiological parameter." For the purposes of this project, the definition of stress put forth by Anisman and Zacharko will be accepted. Event-related brain potentials will be used as our "favorite physiological parameter."

Involvement of the Central Nervous System in Stress

While it continues to be rather difficult to define stress from a clinical point of view, there seems to be increasing acknowledgement among investigators that the central nervous system plays the key role in the generation, perception, mediation, and control of stress. "Any way one looks at it, though, the initial stressor must be viewed as having neuronal consequences..." (Anisman & Zacharko, 1982, p. 125). Heninger (1982, p. 107) states that "behavior is a consequence of nervous activity; thus, behavioral attempts to cope with stress are only one visible aspect of a large number of adaptive mechanisms in

response to stress." Pribram and McGinness (1982, p. 497) feel that "hippocampus is seen as playing a critical role in the pituitary/adrenal 'stress' system...". In a review of the biobehavioral science research related to stress, Hamburg and Elliot (1981) state:

> The catecholamines, which are found both in the adrenal and in several parts of the brain, have long been associated with stress . . . Another clearly relevant group of compounds are the endorphins, which are endogenous, morphine-like peptides that are probably involved in brain regulation of the perception of an response to pain. Also of interest are recent studies which indicate that the brain may influence immune function . . . Studies suggest that stressors are risk factors for a variety of infections . . . (p. 417)

Cohen (1982, p. 279) points out that "it is increasingly evident that an extensive network of central nervous system, autonomic nervous system, endocrine, neuroregulator, opioid peptide, and immunologic responses may be involved." All of these authors cite many references to substantiate their claims.

Brain Potentials and Brain Function

As evidenced above, there is a large and growing body of literature to support the concept that the brain is intimately involved in the generation, perception, mediation, and control of stress. Since EP/ERP and EF/ERF measurements reflect the state of the electrical potentials and magnetic fields of the brain respectively, perhaps they can be used to either detect or further analyze stress in humans. Shagass (1972c, p. 111) states that "our basic assumption is that disordered behavior is associated with altered cerebral excitability and that some aspects of these excitability changes will be reflected in evoked responses." There is some support for this approach.

Both EP/ERPs and EF/ERFs are preceded by either the propagation of action potentials or the spread of a graded potential (Kaufman & Williamson, 1982). In either case, an intracellular axial current along the length of the involved portion of the neuron results. In the case of EP/ERPs, the potential being measured is related to the extracellular volume current that subsequently follows. EF/ERFs, on the other hand, may reflect the intracellular axial current. Both methods would be expected to mirror changes in the activity of large focal populations of neurons and, hence, be correlated with focal brain activity. Indeed, when evoked potentials are recorded over an appropriate area of the sensory cortex, one sees a clearcut response to the sensory stimulus (Desmedt, 1979). Likewise, when the stimulus is internal in origin, consisting of information from past experiences or anticipation in preparation for decision making, significant changes also take place in the event-related potentials. While it is relatively easy to detect these changes, it is enormously difficult to analyze them as to their actual brain mechanism. Determining exactly what areas of the brain are involved at any given moment, what role they serve, and how they relate to and can be modified by all of the various external and internal stimulus parameters remains an elusive goal of research in the area.

Brain Potentials and Stress

Since stress is a psychophysiological state of the brain that involves many functional areas, connecting pathways, and electrical and chemical changes, it is likely to be accompanied by highly complex and variable changes in the EP/ERPs. The problem that can be anticipated is not in being unable to detect a change at all but, rather, in detecting some specific pattern of change that is typical of stress. Hopefully, several different

EP/ERP patterns will be found, each of which can be related to a different type of stress response. From an intuitive point of view, this is a feasible objective, provided one seeks only large, qualitative EP/ERP differences. It would not be reasonable to expect that present knowledge and techniques would allow one to detect bioelectrically that a component of a current EP/ERP response to stress is actually the result of an incident that took place several years ago. On the other hand, it may be possible to detect the difference between the way the brain mediates a calm, effective coping response to stress, on the one hand, and a highly emotional explosion of fear and anxiety on the other hand. In the latter case, a large segment of brain might be intensively activated over a significant period of time; in the former, that area (or areas) might be relatively quiescent.

Few attempts have been made, thus far, to characterize the EP/ERP changes during stress. Shagass (1972a) suggested that the later portion of the ERP may be diminished in amplitude under conditions producing stress or anxiety. McCallum (1979) relates that Sano reported slowly changing potentials during stress. Callaway (1975) speculates about a U-shaped relation between the effects of stress and the amplitude of the later portion of the ERP; as stress intensity increases, so does the ERP amplitude, up to a point. Further increases in the level of stress cause a decline in amplitude. The current situation, then, is one in which a strong need exists for being able to predict the type of reactions likely to be manifested by a given individual when stressed. Considering what is known of stress physiology, the utilization of EP/ERPs and EF/ERFs may provide a practical tool for making such predictions feasible, at least at a rudimentary level.

Predicting Responses to Stress

Individuals vary in their types of responses to stress, and there is some degree of consistency to the type of response of a given individual. It has been suggested that the type of response manifested by a given stressed individual is related to his basic personality (Chesney & Rosenman, 1983; Horowitz, 1976). If the latter relationship is valid, it raises the possibility that EP/ERP and EF/ERF measurements obtained in the baseline state (i.e., while the subject is not being stressed) could be used to predict stress responses indirectly, by virtue of their correlation with classifications of basic personalities. This would, in turn, require that there are correlations between classes of personalities and EP/ERPs.

Personality Types

A number of taxonomies have been introduced to partition personalities into classes that are clinically relevant. Some, like the "Type A and Type B" behavioral patterns (Chesney & Rosenman, 1983), are oriented primarily toward matching personalities with certain patterns of clinical responses to stress. Other, more general schemes have been applied to the problem of enumerating a complete set of elementary personality traits and developing questionnaires to detect reliably the degree to which any of those traits are present in a given individual. One of these, the Eysenck Personality Inventory (EPI) and Questionnaire (EPQ) (EDITS, 1975) has specifically addressed the problems of correlating personality trait measurements with reactions to stress. The "neuroticism factor" of the EPI is described as a measurable variable that "implies low tolerance for stress whether it be physical as in painful situations, or psychological as in conflict or 'frustration' situations" (Eysenck, 1967, p. 41).

There is good reason to expect good correlations between stress reactions and the results of personality tests. As pointed out previously, certain kinds of people do seem to

react in correspondingly stereotyped ways to stress. Further, it is generally acknowledged that stress responses are shaped by "physiological status, genetic traits, current expectations, past experiences.." (Hamburg & Elliott, 1981, p. 414); yet each of these is a vital determining factor of what is called the personality.

Note that personality questionnaires assess the tendency to respond to a stressful situation in a general way. They do not answer the question as to whether or not the security guard will draw his revolver and shoot the intruder. They do attempt to predict whether there is a strong likelihood that, on the one hand, the guard will panic and "fall apart" or, on the other hand, will maintain control and manifest attempts to cope with the situation on the basis of previous and current experiences. The "N" (neuroticism) factor of the EPI is reported to be useful in this regard (Eysenck & Eysenck, 1969). Eysenck (1967) was careful to point out that there are:

> complex interactions between amount of drive present, task difficulty, stress experience, and the various other independent variables . . . Proper quantification of all these variables is essential before confident predictions can be made in the individual case. Without such quantification the theory may still be useful in predicting performance at extreme ends of the scale . . . . (p. 52)

Of course, such tests cannot be repeated too often without becoming ineffective, and they are more or less susceptible to either deliberate or inadvertent misrepresentations on the part of the subject. Further, they require at least a minimum of cooperation, communication, and understanding on the part of the subject. This suggests that they would not be as useful as one might hope for predicting stress responses. As will be outlined later, personality tests can, however, be employed as an independent measure within a battery of measurements needed for prediction.

Personality and Brain Potentials

If one accepts that there is a relationship between (1) personality traits and individual responses to stressful situations, (2) individual responses to stressful situations and the state of that portion of the central nervous system (CNS) that embodies the response to stress, and (3) that portion of the CNS that embodies the response to stress and EP/ERP waveforms, then correlations between personality traits (as revealed, for example, on the EPI) and EP/ERPs should be detectable.

O'Connor (1980) found small differences between subjects with introvert and extravert personalities in amplitudes and sites of origin of their EPs. Shagass (1972a) found a U-shaped relationship between evoked potential amplitude and age with greater amplitudes in childhood and after age 40. He speculated that this relationship behaves differently in extraverts as opposed to introverts in that extraverts may manifest amplitudes, corresponding to those of introverts, at older ages. Friedman and Meares (1979) stated that extraverts show larger amplitudes of the late components of the auditory-evoked potentials, as opposed to the findings of Stelmack, Achorn, and Michaud (1977), which showed greater amplitudes in introverts.

These studies indicate that further explorations of the relationships between personality traits and static EP/ERP patterns might prove fruitful (Eysenck, 1967, p. 261) in the search for predictors of responses to stress. There are, however, several reasons for preferentially pursuing another approach. To begin with, the subject's response to a personality questionnaire may not be honest and/or valid, as outlined previously. Furthermore, the extent to which the EPI neuroticism factor is an indicator of a stress-response

trait is not at all clear, since the neuroticism factor itself may be significantly influenced by the current state of the subject's level of stress (Eysenck & Eysenck, 1969; Hare, Payne, Laurence, & Tawnsley, 1972) and thus reflect the influence of both genotypic and environmental state components.

Baseline EP/ERP measurements alone cannot suffice in attempting to measure/predict stress, even when supplemented by personality assessment questionnaires (McGrath, 1970c). Instead, EP/ERP measurements must also be obtained, for comparison, while the subject is being actively stressed. This view conforms with that of Eysenck (1967), who pointed out that:

> Different reactor systems do not necessarily react in similar ways to stress . . . some measures show significant differences between normals and neurotics during rest, anticipation, stress, and poststress periods. Some measures show differences during all stages except under stress; yet other measures show differences only during stress, and . . . others again yield their main differences during post stress periods. (p. 73)

Further support for this kind of approach consists of experimental neurochemical studies showing that, during stress, selected CNS neurons can be transiently activated with associated chemical changes of the kind that would be expected to affect EP/ERP patterns (Anisman & Zacharko, 1982). It is, therefore, possible that only during stress will there occur specific EP/ERP changes in a subject that can be used to predict the subject's behavioral responses to future stress.

Methods for Experimentally Provoking Stress

This section examines the ways in which someone might be subjected to experimentally induced stress.

Physical and psychological stressors are so diverse in nature that it is hard to imagine any single stressor that would be representative in affecting all subjects in the same way or to the same degree. Moreover, the more effective stressors tend to be those that some might consider to be unethical, such as stress protocols involving the use of electric shocks or deliberately misleading and threatening statements. An acceptable stressor task, for our purposes, is one that (1) can cause no harm, (2) can effectively and uniformly stress all subjects, (3) can be implemented quickly, (4) has short-lived effects, (5) can be applied in a laboratory environment where the subject has little freedom to move, (6) can be used repeatedly with the same effects each time, (7) will not interfere with bioelectrical and biomagnetic recording, and (8) resembles the psychological effect we seek to study (as opposed to more mechanical-physiological "stressors" such as heat, cold-pressors, etc.). Since no known experimental stressor currently fulfills all these criteria, any work of this kind involves using the most reasonable compromise that fits the demands of the experimental situation. However, being prepared to compromise still does not resolve the dilemma. For example, in a statement that a stressor "effectively and uniformly stresses all subjects," what is meant by the word "stresses"? From a practical point of view, what exactly is the stressor supposed to do? How does it achieve its objective? Should it cause a subjective feeling of some sort in the subject, a degradation in the performance of some task, or a change in some physiological measurement? Should it do these things by overloading the subject with a demanding task that is impossible for him to complete, by pain, by exposure to "frightening" scenes, or by conflicting and inconsistent tasks?

Stressors: The Role of Arousal. As one might expect, there is no agreement in the literature as to what constitutes a "stressor," since there is no generally accepted operational definition of what constitutes "stress." As previously noted, some believe that stress and arousal are equivalent (Mason, 1975). If this view of stress as equivalent to arousal is true, then the implication is that a stressor (and only a stressor) must always cause arousal. Others regard arousal as one of the consequences of stress or, in fact, as an essential component thereof (Gray, 1982). This view, in turn, is rebutted by the increasing acceptance of boredom and monotony as stressors (McGrath, 1970c; Appley & Trumbull, 1967); for example, in the case of security guards. Also, this view makes it difficult to explain those cases where chronic severe stress continues beyond the point of exhaustion, at which point fatigue, depression, and inactivation dominate the picture as opposed to arousal (Sanders, 1983). It seems clear that the association between arousal and stress is intimate and strong, but it is not universal; therefore, one cannot legitimately use the level of arousal, or any other effect of arousal, as a measure of stress (Cohen, 1967). The implication of this to stress studies using EP/ERP measurements is obvious--we must clearly differentiate between EP/ERP changes as a result of a stressor's ability to evoke arousal as opposed to its ability to evoke stress. Conversely, any study purporting to measure the effects of a stressor must closely monitor the subject's level of consciousness (arousal), as well as other variables, as possible sources of contamination.

Johnson and Lubin (1972) pointed out that "one can only guess at the number of studies that have been done using subjects who were supposed to be awake but actually dozed off or even slept through the experiment." They emphasized the need to control for changes in level of consciousness as measured by the EEG. Shagass (1972b, 1972c) discussed the effect upon the evoked potentials of alterations in the state of awareness, as correlated with EEG changes in both amplitude and frequency. So did Aleksandrova (1972, p. 107), who concluded that "the higher the alpha rhythm frequency, the shorter the latent period of nearly all EP components in the occipital and vertex regions" and "at a greater alpha rhythm amplitude, longer latent periods of EP components in the vertex region are observed." An exploratory study in the Center's laboratory using continuous EEG recording during evoked potential studies confirmed that there are marked and rapid fluctuations in the level of consciousness of the subjects that may not be readily apparent to either the subject or the technician obtaining the EP data. In short, stress cannot be regarded as equivalent to arousal; hence, a stressor cannot be defined as any stimulus that produces arousal, nor can the level of stress be measured by the level of arousal. On the contrary, in any experimental study of stress, an attempt must be made to monitor and maintain a constant level of arousal.

Detecting and Monitoring Stress. Since there is no agreement on the definition of stress according to its intrinsic mechanisms, perhaps stress can be defined as a pattern of physiologic responses. Although this approach would afford the advantage of being able to select or evaluate a potential stressor by virtue of its facility in evoking those responses, it requires selection of a proper set of responses. Activation of the adrenocortical endocrine system is not specific for stress (Mason, 1975). Autonomic nervous system activation is well known to be unreliable, to the point that others have used this unreliability as evidence of the existence of a high degree of individual specificity and intersubject variability in responding to the same stressor (Lacey, Kagan, Lacey, & Moss, 1963). Performance measures likewise are characterized as not being proper indicators for stress (Sanders, 1983). Sanders, like McGrath (1970b), has recommended that performance measures should be used only as control measures to ascertain that sufficient effort is allocated to keep performance at the optimum during the task. Indeed, if we believe that the CNS plays a role in stress and therefore in using EP/ERP measurements

to assess stress, we are placing ourselves in the curious stance of trying to select and us other physiologic or performance methods, already known to be relatively unsuccessful a detecting or measuring stress, as indicators by which to judge the efficacy of a technique suspected to be far more sensitive and selective.

Without some other reliable method of measuring stress, it will not be possible to assess the significance of an EP/ERP measurement that shows no change during a stressor test. Finding no change could mean that the stressor used was not, after all, effective. It could also mean that EP/ERP measurements are not capable of detecting changes in brain activity during stress. Finally, it could mean that the stressor is effective on most subjects but, for some reason, not on this particular subject. This last possibility implies that there may be general classes or types of stressors, and that there may be relationships between the subjects who show EP/ERP responses to a particular type of stressor and their basic personalities. For example, extraverts might tend to manifest a significant change in response to any one of a whole class of stressors, while introverts would show no change to those same stressors.

Types of Stressors. There have been a number of publications that seek to subdivide experimental stressors into classes based primarily upon protocols; for example, upon whether or not the stressor task involves time-sequencing of stimuli, the interpretation of complex information, or threats of punishment (Hackman, 1970). Little, if anything, has been published that attempts to relate classes of stressors to types of responses or personalities, except for the previously mentioned work regarding the autonomic nervous system and individual differences, as with Type A/B behavior profiles. This lack is curious, from a clinical point of view, because the existence of such classes of stressors and related classes of responders is an everyday, and often dramatic, observation. Witness the person who is "self-motivated" to the extent that he is severely stressed by his own internal demands upon his performance, compared with another person who becomes stressed mainly by more primitive external imagery and could not care less about time and task performance. The former personality often characterizes those who develop duodenal ulcers (Alexander, 1950), while the latter often seems to be the case in those who possess hysterical personalities (Horowitz, 1976).

Protocols. The implication of these observations is that it should be possible to better evaluate the effectiveness of, and response to, a particular stressor by using several stressors on each tested subject, with each stressor designed to best evoke stress in a subgroup (of the general population) characterized by a specific personality profile. This approach is not a new concept in the stress literature. Although the need for studies of this sort has often been pointed out (McGrath, 1970d), seldom have they been done. Basically, the voluminous literature on stress tasks recommended that the stressor task protocol should:

1. Include baseline evaluations of the subject.

2. Include measurements of personality assessments.

3. Include evaluations of the response to several different types of stressors.

4. Use each given type of stressor at multiple levels of intensity.

5. Include ongoing assessments of arousal, particularly EEG recording.

6. Measure responses by means of several different parameters.

7. If possible, be repeated over a period of time, on the same subjects.

8. Be implemented in the most meticulously controlled environment as is reasonable for the test and the objectives of the study.

9. Consume a reasonably short period of time so as to avoid marked changes in level of consciousness and in other physiological variables.

Analysis of Data

The manner of data interpretation warrants further discussion. Recall that EP/ERP data are obtained by averaging together many observations. This must be done because the bioelectric activity of interest is of very low amplitude and is often obscured by the background activity. The averaging reduces the apparent effect of the background because the background frequency is not time-locked or synchronous with the stimulus. The evoked potential, on the other hand, is synchronized to the stimulus and thereby enhanced by averaging. It is believed that the earlier components of the evoked potential (up to 100 msec) largely represent primary sensory processing (Hillyard, Picton, & Regan, 1978) and that the later components represent "cognition"; that is, the higher-level central nervous system handling of the input (e.g., various aspects of recognition, association, storage, coordination, evaluation, and other functions that are applied to or affected by the stimulus) (McCallum, 1979).

Investigators often obtain such evoked potentials, both under the "nominal" laboratory situation (baseline recordings) and during the experimental stress condition. The data are then analyzed for differences between the waveform amplitudes and latencies of the two recordings, and any differences are attributed to the effect of the situation (e.g., stress) upon the individual. This approach may have both empirical and theoretical shortcomings.

Note that, in this approach, one does not look generally at the manifestation of stress within the brain. Rather, one is looking selectively at the effect of the resultant stress upon the brain's response to a visual or auditory stimulus. This particular approach of analyzing auditory or visual stimuli may have absolutely no clinical or behavioral relevance to stress as an entity worthy of independent study.

While the evoked brain response to the visual or auditory stimulus is time-locked and is therefore enhanced by averaging, the brain's response to stress is not time-locked and is not transient. It is present and changing all the time that the evoked responses are being obtained; therefore, averaging techniques will not necessarily represent meaningful information specifically related to stress. In particular, the "background activity" that is thereby averaged out may be the very same activity that is most meaningful and that should be captured and analyzed (Callaway, 1979). While empirical, statistical explorations of this sort might luckily hit upon some consistent relationship, more direct methods are desirable. Therefore, in addition to the previous recommendations, one should at least explore the feasibility of alternative approaches. One such approach would be to eliminate the visual/auditory stimulus, and, instead, consider the possibility of using a time-locked stimulus consisting of the stressor itself. The stressor would obviously have to be one that has rapid onset, short duration, and is repeatable so as to enable its resulting EP/ERP waveforms to survive and even be enhanced by the technique of averaging. As an example, the stimuli could consist of a sequence of pictures flashed upon a screen at controlled intervals of time. The pictures could be selected so as to evoke various degrees of stressful feelings in the subject. This particular example,

however, would have the disadvantage of also causing visual-evoked cerebral potentials. Finding a suitable stimulus to meet these qualifications is, admittedly, a challenge, but appears to be a promising approach.

Another approach would be to study the responses to stress as manifested i topographical EEG displays/analyses, a technique that allows one to dispense with al stimuli other than the stressor (Livanov, 1977).

Data Displays

The current format for displaying the evoked potential, where amplitude changes ar plotted against time for each of the different recording sites, presents the data in a wa that makes it difficult for the investigator to discern dynamic anatomical-tempora fluctuations in EP/ERP amplitudes. Vaughan (1979) states that:

> There is an intrinsic ambiguity in the interpretation of scalp potential amplitude variations--they may reflect either changes in amount or extent of neural activity, an ambiguity which can be resolved only by detailed mapping of the surface potential distribution.
>
> It is evident, therefore, that quantitative analyses of the ERP must evaluate not only the magnitude and timing of their components, but also their spatial distribution (p. 444).

Further, Vaughan (1982) urged that topographic data be used to interpret the surfac recorded ERP distributions in terms of their intracranial sources. Duffy (1982, p. 19 concurs with Vaughan:

> The interpretation or evaluation of multichannel EP data requires analysis of large volumes of data across both space and time. We propose that the inherent difficulties involved in making such spatio-temporal correlations by unaided visual inspection place constraints on both the clinical utility and research applicability of EP. The topographic mapping system . . . reduces the dimensionality of data and offers . . . major advantages.

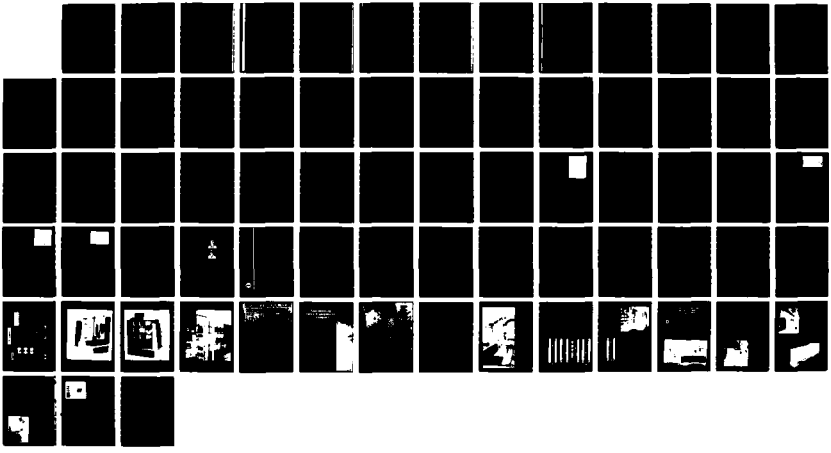
Topographical Displays

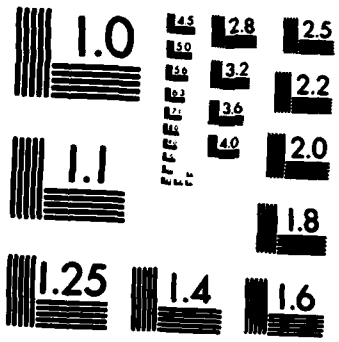
Temporal mapping has been employed for many years by increasing [illegible] investigators (Livanov, 1977; Ragot & Remond, 1979). By using displays wherein [illegible] colors and hues represent different polarities and amplitudes, one can even [illegible] represent and analyze the origin and spread over the surface of the brai[illegible] changes that take place over a period of time following the stimulus. [illegible] been used both with EEG and EP/ERP studies and also have incorporate[illegible] applications to the data before they are displayed (Duffy, 1982).

This approach is referred to as "color topographic[illegible] partitioned in such a way as to represent a diagram of the [illegible] All the cortical regions are thereby represented simultaneously [illegible] different amplitudes of the electric fields over each regi[illegible] colors and hues. Most displays of this sort allow at least 256 [illegible] spread of, for example, a positive-polarity high-ampl[illegible] dynamic, movie-like or cartoon fashion as it first appears [illegible] to a diff[illegible] cortical location in succeeding time inter[illegible]

AD-A188 809 PROCEEDINGS OF THE BRAIN MAPPING MACHINE DESIGN WORKSHOP HELD IN COLLEGE..(U) TEXAS A AND M UNIV COLLEGE STATION R B LIVINGSTON AUG 85 5/5

UNCLASSIFIED DAMD-17-85-G-5042 F/G 6/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

however, would have the disadvantage of also causing visual-evoked cerebral potentials. Finding a suitable stimulus to meet these qualifications is, admittedly, a challenge, but appears to be a promising approach.

Another approach would be to study the responses to stress as manifested i topographical EEG displays/analyses, a technique that allows one to dispense with al stimuli other than the stressor (Livanov, 1977).

Data Displays

The current format for displaying the evoked potential, where amplitude changes ar plotted against time for each of the different recording sites, presents the data in a wa that makes it difficult for the investigator to discern dynamic anatomical-tempora fluctuations in EP/ERP amplitudes. Vaughan (1979) states that:

> There is an intrinsic ambiguity in the interpretation of scalp potential amplitude variations--they may reflect either changes in amount or extent of neural activity, an ambiguity which can be resolved only by detailed mapping of the surface potential distribution.
>
> It is evident, therefore, that quantitative analyses of the ERP must evaluate not only the magnitude and timing of their components, but also their spatial distribution (p. 444).

Further, Vaughan (1982) urged that topographic data be used to interpret the surface recorded ERP distributions in terms of their intracranial sources. Duffy (1982, p. 19 concurs with Vaughan:

> The interpretation or evaluation of multichannel EP data requires analysis of large volumes of data across both space and time. We propose that the inherent difficulties involved in making such spatio-temporal correlations by unaided visual inspection place constraints on both the clinical utility and research applicability of EP. The topographic mapping system . . . reduces the dimensionality of data and offers . . . major advantages.

Topographical Displays

Temporal mapping has been employed for many years by increasing numbers of investigators (Livanov, 1977; Ragot & Remond, 1979). By using displays wherein different colors and hues represent different polarities and amplitudes, one can even more easily represent and analyze the origin and spread over the surface of the brain of EP/ERP changes that take place over a period of time following the stimulus. Such displays have been used both with EEG and EP/ERP studies and also have incorporated statistical applications to the data before they are displayed (Duffy, 1982).

This approach is referred to as "color topographical displays." The screen is partitioned in such a way as to represent a diagram of the cortical surface of the brain. All the cortical regions are thereby represented simultaneously. As time elapses, the different amplitudes of the electric fields over each region are indicated by different colors and hues. Most displays of this sort allow at least 256 different hues, so that the spread of, for example, a positive-polarity high-amplitude peak can be presented in a dynamic, movie-like or cartoon fashion as it first appears in one location and then travels to a different cortical location in succeeding time intervals.

The benefits derived from neurochemical studies are well-known. They are responsible for much of our current knowledge of brain function. The limitations of neurochemical studies, however, are also well-known. It is information ordinarily derived from experimental animal material, autopsies, or invasive procedures. The process of obtaining or analyzing the material often destroys or significantly alters the function of its structures. Further, knowing that a chemical is present at a specific location does not necessarily tell its function there or the function of that part of the brain. It is a static measurement in that it tells little of the moment-to-moment spread of control or data through the brain, and, hence, little of the overall purpose or mechanism of neuronal activities. While a number of different models of brain function have been proferred, even in regard to "anxiety" (Gray, 1982) and "stress" (Anisman & Zacharko, 1982), those models all hypothesize and require a very selective sequence of discharging in groups of neurons, whose effects are transmitted over specific pathways to and from specific brain areas and in a specific timed sequence. Confirmatory evidence for any such model has to come from studies that are noninvasive and dynamic, such as EP/ERP measurements. However, the evidence will be helpful only to the extent that it provides information now lacking. It is necessary to know which brain areas are active, at what time, and in what sequence the information spreads over which specific pathways from one area to the next.

According to Nebylitsyn (1972), efforts in Russia since the time of Pavlov to explain the properties of the central nervous system from either a unified point of view, on the one hand, or a "partial, regional" point of view, on the other hand, were unsuccessful. He postulated that it was not reasonable that brain functions could be subdivided into three or four major regional properties, as Pavlov had anticipated. On the contrary, he perceived the neurophysiological parameters of behavior to be far more complex, composed of a larger number of smaller regions, each of which is primary for some function.

> The level of one or another parameter of the prefrontal cortex may not, for example, coincide functionally with that of the same parameter measured for the medio-basal region of the frontal lobes or for the limbic formations. Factual information relevant to problems of this kind will, however, only be gained by experimental investigations involving the application of methods yet to be worked out, and which will permit us to establish the characteristics of activity in spatially different nervous structures in the regulatory system. (p. 412)

Nebylitsyn found interesting the topographical EEG studies by Livanov (1977) showing that methods of this sort could be used to analyze intellectual operations. Nebylitsyn concluded that:

> This approach, in its turn, opens perspectives for the creation of a psychophysiologically-based system of personnel selection tests and for the elaboration of the psychophysiological aspects of the theory of human reliability in complex working conditions (p. 414).

Indeed, Livanov (1977) was able to differentiate individuals, according to the difficulty they encountered in solving mentally an arithmetic problem, by virtue of run-time topographical analysis.

Cognitive and emotional activities are relatively long events. They are undoubtedly highly complex, involving repeated interactions among many separate parts of the brain in a complicated sequence of events. It is, therefore, highly unlikely that one can ever begin to understand the p[illegible] by looking at the activity of only one part of the brain at one

time. On the contrary, a method must be used that presents the information in a w that answers can be found to questions such as those listed below:

1. Where in the brain does the first visible response to stress (in this person) to appear?

2. Where does it go from there?

3. Does the pattern differ among subjects?

4. Are there groupings of subjects possessing similar patterns?

5. Do the patterns differ regarding the immediate effects of stress or the reac to stress?

6. Do those groupings correspond to personality assessment and/or beh responses?

7. Are there patterns consistent with any of our models of stress?

Simply looking at a large numerical tabulation of amplitudes or latencies wil readily accomplish this, nor will statistical compilation performed upon two or "curves" or visual "curve" inspections. Color topographical displays, on the other are well suited to this purpose.

At this time, color topographical analysis cannot be applied to data from EF (magnetic field) measurements and, therefore, can be used only in EP/ERP measurem This is because the magnetic sensors are large and cumbersome, using liquid helium cooling purposes. They do not allow multiple independent probes to be used simul ously on the subject. The likelihood exists that this limitation will be overcome i next few years. Meanwhile, studies on stress of the sort with which we are concerne best confined to EP/ERP measurements. Separate studies may continue on the bio netic-bioelectric comparisons and on the mechanisms of the origin of biomag activity, until more suitable equipment is developed. At that point in time, topographical techniques can be easily adapted to this newer technology.

Testing for Stress, Stressors, and Individual Differences

Preliminary Work

A small-sample prototype of a stress-task protocol was conducted on three sub at the Center during the summer of 1981. It was done, realizing the almost unlim criticisms to which it would be susceptible, simply to explore the feasibility and pro of that approach in conjunction with color topographical displays. Considerable straints were imposed at that time upon the protocol as the result of the then-cur computer hardware, which lacked sufficient main memory and secondary storage capa to accommodate the large amount of data that needed to be manipulated. Nonethe the displays were found to be easily and quickly interpreted. The stressor was effectiv causing marked changes in the pattern of EP/ERPs across the scalp. Differences in t patterns among the subjects were profound. It would be premature to interpret results further, except to conclude that this kind of approach is not only feasible but highly promising.

The Future

Current theoretical models of brain function, while of great interest, seem to have reached an impasse characterized by contradictions and a search for solutions that only new data can supply. For example, Riss (1983), who developed a computer model of human visual processing, feels that magnetoencephalography represents the only noninvasive tool now available to clarify the sequence of activation of the different foci he has identified as a part of the visual process. Also, Zuckerman (1982) stated that studies "looking for the biological correlates of the psychological variable . . . can only approach causation through techniques such as path analysis . . ." The derivation of topographical information from EP/ERP/EEG studies of stress represents the kind of approach that can be adapted to the study of visual processes, as well as other brain functions, particularly those of cognitive nature. The application of this approach will hopefully lead to a greater understanding of brain functions generally and also as they vary from individual to individual under specific circumstances. While this largely empirical approach can be expected to yield significant new and pertinent information about the brain, personality, and behavior, one should not neglect efforts that are directed toward understanding conceptually how the brain might handle "information processing" in circumstances such as stress. In particular, computer simulation and emulation models, expert systems, and other theoretical endeavors in the realm of artificial intelligence (AI) can be quite useful when developed in conjunction with information learned from topographical EP/ERP/EEG stress protocol approaches. They can provide the framework that will enable researchers to better understand and explain empirically derived data and to guide them in further applying such tools as color topography. Newell (1983) has expressed this well:

> AI would appear to be at the mercy of the immense gulf that continues to separate psychology and the biology of the brain. As each field continues to progress--which both do dramatically--hopes continually spring up for new bridging connections. No doubt at some point the permanent bridge will get built. So far, although each increment of progress seems real, the gap remains disappointingly large.
>
> It is possible that AI has a major contribution to make to this by exploring basic computational structures at a level that makes contact with neural systems.

A high-speed, 32-bit, large-capacity, stand-alone computer will become available for use within several months. With the anticipated acquisition of a color monitor and suitable software, the Center's laboratory will be in a position of being able to implement a well controlled study dealing with the ability of EP/ERP topographical analyses to predict future job performance under stress and to explore the possibility of bridging the gap between psychology (stress) and biology (EPs/ERPs) by using computer-based models.

RECOMMENDATIONS

The following recommendations are made for a research protocol for ascertaining whether measurements of brain activity can be used to predict performance under stress.

1. Ideally, subjects should be in residence at the testing center from at least the prior evening to maintain some control over the environment. Too often, subjects violate the rules regarding sleep, food ingestion, etc. Since this cannot be implemented at NAVPERSRANDCEN, greater effort must be placed upon emphasizing the importance of

any restrictions in these areas and adequately interviewing subjects about adherence to them prior to testing.

2. Personality testing, using the EPI/EPQ should be done, preferably on a separate day prior to testing.

3. Subjects should be available for repeated testing at appropriate intervals. Considerable efforts should be made to obtain a thoroughly representative group of subjects. At the very least, sufficient personal information must be obtained so that the subject population can be characterized in detail (i.e., gender, age, work, diet, medications, health, mental state, visual and auditory acuity, and so on).

4. Recording sessions should involve the following factors:

a. Multiple electrodes (8 channels at the very least) must be used.

b. Both bipolar and monopolar recording should be considered. Monopolar recording provides a more accurate reflection of the local amplitudes and frequencies whereas bipolar recording facilitates the recognition of extracerebral artifacts and local activity relative to other cerebral locations. Some further exploratory studies might be of benefit regarding the selection of specific reference leads and montages. The literature seems to favor monopolar references to the ear leads.

c. Ongoing, continuous EEG recording must be used to assess the subject's state of alertness (refer to the previous discussion on stress vs. arousal).

d. The subject should be in a comfortable sitting position and be vigorously alerted between all runs, with lights turned on, conversation, novel noises, etc. The room must be kept cool to facilitate alertness. The technician should make notes, in conjunction with time-markers during the recordings, with reference to the subject's level of alertness.

e. Consideration should be given to checkerboard-reversal stimuli rather than light flashes or static checkerboards since this seems to be more and more the "industry standard" in the effort to achieve reproducible waveforms (Starr, Sohmer, & Celesia, 1978).

f. Continuous EKG (heart rate) monitoring should be done throughout the testing so as to have available an additional and independent measurement of stress, even though not fully reliable (Kak, 1981).

g. Testing should be preceded by revisions to existent computer software so as to allow more rapid employment of the testing procedures.

h. The test should be partitioned as follows:

(1) Baseline recording, followed by Stressor #1 recordings:

(a) Low intensity followed by a brief rest interval.
(b) Medium intensity followed by a brief rest interval.
(c) High intensity followed by a brief rest interval.

(2) Baseline recording, followed by Stressor #2 recordings:

(a) Low intensity followed by a brief rest interval.
(b) Medium intensity followed by a brief rest interval.
(c) High intensity followed by a brief rest interval.

i. Stressor #1 should be a timed, cognitive, performance task requiring little, if any, subject movement. The cube-counting task is ideal here, particularly since this laboratory has conducted, over the past few years, a large number of studies that have incorporated an EP/ERP cube-counting performance task. During this task, the subjects are presented with drawings of various numbers and configurations of three-dimensional stacks of cubes. They are instructed to count accurately the number of cubes represented in each drawing within a fixed amount of time. Successive drawings are more complex and therefore more difficult and "stressing" to count within the required amount of time. The resulting cube-count scores can be used to assess performance effort, while simultaneously obtained heart rate and EP/ERP data can be used to measure and analyze the response to stress. The procedure is amenable to the use of minimal deception to enhance the stress; for example, by informing the subject that performance thus far is poor. This task requires little motor movement, provides ways of assessing performance effort, allows for multiple levels of difficulty, is easily repeatable, is not harmful to the subject, is easily reproduced from subject to subject, does not interfere with auditory-evoked potential studies, consumes a short period of time, and would be expected to cause stress in a subset of the population (i.e., those who are self-motivated and highly conscientious individuals), thereby inviting correlation studies regarding personality traits. The Stroop test, and others of this sort, could also be considered. Here, as one example, the subject is shown, on a screen, the name of a color. The screen image is also colored, but the color of the screen image does not necessarily match that of the name. The subject's task is to select one out of a set of objects where the object's color coincides with that of the name on the screen but not necessarily that of the image's color. The Stroop test may also be timed and shares many of the advantages of the cube-counting task. Stressor #2 should be of a completely different nature, such as pictorial presentations of potentially disturbing photographs intermixed with neutral scenes.

j. Magnetic studies should be done separately so as not to introduce further constraints on the stress protocol methods.

k. The testing should employ summated auditory-evoked potentials as targets for analysis, since this can be implemented with little disturbance of conscious effort upon task performance.

l. Analysis of the data should include the technique of color topography.

m. Subjects should be interviewed (or fill out a questionnaire as previously developed at the Center) after testing, in order to independently assess the subjective attitudes and degree of stress during the testing procedures.

n. A separate protocol should be developed and implemented wherein evoked potentials are not used. Instead, the stressor itself should be the time-locked stimulus and either part of a single epoch study or used as part of a sequence of stressors for summations. Pictorial displays as alluded to above could be adapted for this. Perhaps this technique of equating the stressor with the time-locked stimulus will, in the end, prove to be more meaningful and useful than any other.

REFERENCES

Aleksandrova, N. I. The correlation between background alpha activity and the characteristics of the components of evoked potentials. In V. D. Nebylitsyn & J. A. Gray (Eds.). Biological bases of individual behavior. New York: Academic Press, 86-110, 1972.

Alexander, F. Psychosomatic medicine: its principles and applications. New York: W. W. Norton & Co., Inc., 1950.

Anisman, H., & Zacharko, R. M. Depression: The predisposing influence of stress. Behavior and brain science, 1982, 5, 89-152.

Appley, M. H., & Trumbull, R. On the concept of psychological stress. In M. H. Appley & R. Trumbull (Eds.). Psychological stress: Issues in research. New York: Appleton-Century-Crofts, 1967, 1-13.

Begleiter, H., & Porjesz, B. Evoked brain potentials as indicators of decision-making. Science, 1975, 187, 754-755.

Callaway, E. Brain electrical potentials and individual psychological differences. New York: Grune and Stratton, 1975.

Callaway, E. Individual psychological differences and evoked potential variability. In J. E. Desmedt (Ed.). Cognitive components in cerebral event-related potentials and selective attention. New York: S. Karger, 1979, 6, 243-257.

Chesney, M. A., & Rosenman, R. H. Specificity in stress models: Examples drawn from Type A behavior. In C. L. Cooper (Ed.). Stress research. New York: John Wiley & Sons, Ltd., 1983, 121-146.

Cohen, F. Stress and bodily illness. Psychology clinics of North America, 1982, 4(2), 269-286.

Cohen, S. I. Central nervous system functioning in altered sensory environments. In M. H. Appley & R. Trumbull (Eds.). Psychological stress: Issues in research. New York: Appleton-Century-Crofts, 1967, 1-13.

Cohen, S. Aftereffects of stress on human performance and social behavior: A review of research and theory. Psychology Bulletin, 1980, 88(1), 82-108.

Desmedt, J. E. Somato sensory evoked potentials in man: Maturation, cognitive parameters and clinical uses in neurological disorders. In D. Lehmann & E. Callaway (Eds.), Human evoked potentials: Applications and problems. New York: Plenum Press, 1979, 83-103.

Donchin, E., Ritter, W., & McCallum, W. C. Cognitive psychophysiology: The endogenous potentials in man. In E. Callaway, P. Tueting, & S. H. Koslow (Eds.). Event-related brain potentials in man. New York: Academy Press, 1978, 223-321.

Duffy, F. H. Topographic displays of evoked potentials: Clinical application of brain electrical activity mapping (BEAM). In I. Bodis-Wollner (Ed.). Evoked potentials. 1982, 388, 183-196.

EDITS Manual: Eysenck Personality Questionnaire (Junior and Adult). 1975. Education and Industrial Testing Service, San Diego.

Eysenck, H. J. The biological basis of personality. Springfield, IL: Charles C. Thomas 1967.

Eysenck, H. J., & Eysenck, S. B. G. Personality structure and measurement. San Diego: Robert R. Knapp, 1969.

Friedman, J., & Meares, R. Cortical evoked potentials and extraversion. Psychosomatic Medicine, 1979, 41(4), 279-286.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: John Wiley, 1966.

Gordon, H. S., Silverberg-Shalev, R., & Czernilas, J. Hemispheric asymmetry in fighter and helicopter pilots. Acta Psychologica, 1982, 52, 33-40.

Gray, J. A. Precis of the neuropsychology of the brain: An inquiry into the functions of the septo-hippocampal system. Behavioral Brain Science, 1982, 5, 469-534.

Grings, W. W., & Dawson, M. E. Emotions and bodily responses: A psychophysical approach. New York: Academic Press, 1978.

Hackman, J. R. Tasks and task performance in research on stress. In J. E. McGrath (Ed.). Social and psychological factors in stress. New York: Holt, Rinehart, and Winston. 1970, 202-237.

Hamburg, D. A., & Elliott, G. R. Biobehavioral sciences: An emerging research agenda. Psychological clinics of North America, 1981, 4(2), 407-421.

Hare, E. H., Payne, H., Laurence, K. M., & Tawnsley, K. Effect of severe stress on the Maudsley Personality Inventory Score in normal subjects. British Journal of Social and Clinical Psychology, 1972, 11(4), 353-358.

Haythorn, W. W., & Altman, I. Personality factors in isolated experiments. In M. H. Appley & R. Trumbull (Eds.). Psychological stress: Issues in research. New York: Appleton-Century-Crofts, 1967, 363-399.

Heninger, G. R. Monoaimine receptor sensitivity and antidepressants. Commentary to H. Anisman and R. M. Zacharko, Depression: The predisposing influence of stress. Behavior and Brain Science, 1982, 5, 107-108.

Hillyard, S. A., Picton, T. W., & Regan, D. Sensation, perception, and attention: Analysis using ERPs. In E. Callaway, P. Tueting, & S. H. Koslow (Eds.). Event-related brain potentials in man. New York: Academic Press, 1978, 223-321.

Hokanson, J. E. The physiological bases of motivation. New York: John Wiley & Sons, Ltd., 1969.

Horowitz, M. J. Stress response syndromes. New York: Jason Aronson, Inc., 1976.

Kak, A. V. Stress: An analysis of physiological assessment choices. In G. Salvendy & M. J. Smith (Eds.). Machine pacing and occupational stress. London: Taylor & Francis, Ltd., 1981, 135-142.

Kaufman, L., & Williamson, S. J. Magnetic location of cortical activity. In I. Bodis-Wollner (Ed.). Evoked potentials, 1982, 388, 197-213.

Kinsbourne, M. (Ed.) Asymmetrical function of the brain. New York: Cambridge University Press, 1978.

Johnson, L. C., & Lubin, A. On planning psychophysiological experiments: Design, measurement, and analysis. In N. S. Greenfield & R. A. Sternbach (Eds.). Handbook of Psychophysiology. New York: Holt, Rinehart, and Winston, Inc., 1972, 125-158.

Lacey, J. I., Kagan, J., Lacey, B. C., & Moss, H. A. The visceral level: Situational determinants and behavioral correlates of autonomic response patterns. In P. H. Knapp (Ed.). Expression of the emotions in man. New York: Internist University Press, Inc., 1963, 161-196.

Lewis, G. W. Event-related brain electrical and magnetic activity: Toward predicting on-job performance. International Journal of Neuroscience, 1983, 18, 159-182. (a)

Lewis, G. W. Bioelectric predictors of personnel performance: A review of relevant research at the Navy Personnel Research and Development Center (NPRDC Tech. Rep. 84-3). San Diego: Navy Personnel Research and Development Center, November 1983. (b)

Livanov, M. N. Spatial organization of cerebral processes. New York: John Wiley and Sons, 1977.

Mason, J. W. A historical view of the stress field, Part II. Journal of Human Stress, 1975, 1, 22-36.

McCallum, W. C. Cognitive aspects of slow potential changes. In J. E. Desmedt (Ed.). Cognitive components in cerebral event-related potentials and selective attention. New York: S. Karger, 1979, 151-171.

McGrath, J. E. Major methodological issues. In J. E. McGrath (Ed.). Social and psychological factors in stress. New York: Holt, Rinehart, and Winston, 1970, 41-57. (a)

McGrath, J. E. Major substantive issues: Time, setting, and the coping process. In J. E. McGrath (Ed.). Social and psychological factors in stress. New York: Holt, Rinehart, and Winston, 1970, 22-40. (b)

McGrath, J. E. Settings, measures and themes: An integrative review of some research on social-psychological factors in stress. In J. E. McGrath (Ed.). Social and psychological factors in stress. New York: Holt, Rinehart, and Winston, 1970, 58-96. (c)

McGrath, J. E. Some strategic consideration for future research on social-psychological stress. In J. E. McGrath (Ed.). Social and psychological factors in stress. New York: Holt, Rinehart, and Winston, 1970, 348-352.

Murison, R., & Ursin, H. Stress as activation. Commentary to H. Anisman, & R. M. Zacharko. Depression: The predisposing influence of stress. Behavior and Brain Science, 1982, 5, 89-152.

Nebylitsyn, V. D. The problem of general and partial properties of the nervous system. In V. D. Nebylitsyn & J. A. Gray (Eds.). Biological bases of individual behavior. New York: Academic Press, 1972, 400-417.

Newell, A. Intellectual issues in the history of artificial intelligence. In F. Machlup & U. Mansfield (Eds.). The study of information: Interdisciplinary messages. New York: John Wiley & Sons, 1983.

Nugent, W. A. Biotechnology predictors of physical security personnel performance: II. Survey of experimental procedures to assess performance under stress (NPRDC Spec. Rep. 84-9). San Diego: Navy Personnel Research and Development Center, November 1983.

O'Connor, K. Electrocortical positivity and personality. Perceptual and motor skills, 1980, 51, 924-926.

Pepitone, A. Self, social environment, and stress. In M. H. Appley & R. Trumbull (Eds.). Psychological stress: Issues in research. New York: Appleton-Century-Crofts, 1967, 182-208.

Pribram, K. H., & McGuinness, D. The anatomy of anxiety? Commentary to J. A. Gray. Precis of the neuropsychology of the brain: An inquiry into the functions of the septo-hippocampal system. Behavior and Brain Science, 1982, 5, 496-498.

Ragot, R., & Remond, A. Event-related scalp potentials during a binaural choice R. T. task: Topography and interhemispheric relations. In D. Lehmann & E. Callaway (Eds.). Human evoked potentials: Applications and problems. New York: Plenum Press, 1979, 303-316.

Riss, W. Testing a theory of brain function by computer methods. Brain behavior evolution, 1983, 22, 42-52.

Robinson, E. R. N. Biotechnology predictors of physical security personnel performance: I. A review of the stress literature related to performance (NPRDC Tech. Note 83-9). San Diego: Navy Personnel Research and Development Center, June 1983. (AD-A131 133)

Sanders, A. F. Towards a model of stress and human performance. Acta Psychologica. 1983, 53, 61-97.

Selye, H. The stress concept: Past, present and future. In C. L. Cooper (Ed.). Stress research. New York: John Wiley & Sons, Ltd., 1973, 1-20.

Selye, H. Stress without distress. Philadelphia: J. B. Lippincott Co., 1974.

Selye, H. Confusion and controversy in the stress field. Journal of Human Stress. 1975, 1, 37-44.

Shagass, C. Evoked brain potentials in psychiatry. New York: Plenum Press, 1972. (a)

Shagass, C. Electrical activity of the brain. In N. S. Greenfield & R. A. Sternbach (Eds.). Handbook of Psychophysiology. New York: Holt, Rinehart, and Winston, Inc., 1972, 263-328. (b)

Shagass, C. Cerebral evoked responses and personality. In V. D. Nebylitsyn & J. A. Gray (Eds.). Biological bases of individual behavior. New York: Academic Press, 1972, 111-127. (c)

Starr, A., Sohmer, H., & Celesia, G. G. Some applications of evoked potentials to patients with neurological and sensory impairment. In E. Callaway, P. Tueting, & S. H. Koslow (Eds.). Event-related potentials in man. New York: Academic Press, 1978, 155-196.

Stelmack, R. M., Achorn, E., & Michaud, A. Extraversion and individual differences in auditory evoked response. Psychophysiology, 1977, 14(4), 368-374.

Trumbull, R., & Appley, M. H. Some pervading issues. In M. H. Appley & R. Trumbull (Eds.). Psychological stress: Issues in research. New York: Appleton-Century-Crofts, 1967, 400-412.

Ursin, H. Activation, coping, and psychosomatics. In H. S. Ursin, E. Baade, & S. Levine (Eds.). Psychobiology of stress: A study of coping men. New York: Academic Press, 1978, 201-228.

Vaughan, H. G., Jr. A neurophysiology of mind? In H. Begleiter (Ed.). Evoked brain potentials and behavior. New York: Plenum Press, 1979, 437-446.

Vaughan, H. G., Jr. The neural origins of human event-related potentials. In I. Bodis-Wollner (Ed.). Evoked potentials, Annals of the New York Academy of Science, 1982, 388, 125-138.

Welford, A. T. Fundamentals of skill. London: Methuen & Co., Ltd., 1968.

Zuckerman, M. Leaping up the phylogenetic scale in explaining anxiety: Perils and probabilities. Commentary to J. A. Gray. Precis of the neuropsychology of the brain: An inquiry into the functions of the septo-hippocampal system. Behavioral Brain Science, 1982, 5, 505-506.

VIETNAM HEAD INJURY STUDY

Department of Clinical Investigation
Walter Reed Army Medical Center
Washington, DC 20307-5001

Andres M. Salazar, COL, MC, USA, Director

INTERIM REPORT

6 March 1985

Veterans Administration
Contract #IGA V101(91) M-79031-2

INTRODUCTION

The Vietnam Head Injury Study (VHIS) registry includes 1221 young veterans who survived penetrating brain wounds from shrapnel or bullets between 1967 and 1970 in the Vietnam War and on whom we have detailed medical records of the initial and follow-up medical care. Phase I of the VHIS, conducted between 1976 and 1979, involved a review and computer codification of these records by experienced neurologists and neurosurgeons and resulted in publication of a number of important scientific papers. Phase II, which was formally begun in 1980 and is still ongoing, involves an extensive, one-week inpatient reevaluation of VHIS registrants who volunteered to be examined and of 85 uninjured Vietnam veterans who volunteered to serve as control subjects. The standardized evaluation included a detailed neurological examination; computerized tomographic (CT) brain scan (which gives the exact size and location of the injury); extensive neuropsychological, behavioral, and speech and language batteries; a physical rehabilitation and motor performance battery; EEG and brain evoked potentials testing; an audiological battery; an extensive social service family interview conducted in the veteran's home by trained American Red Cross personnel; and separate family/community adjustment questionnaires. By the end of the formal evaluation and data collection period of Phase II in October 1984, 520 brain-injured veterans and 85 controls had been evaluated. Over 22,000 data points have been collected on each of these men and computerized for subsequent analysis. The following is a brief summary of analysis completed to date, as well as analyses planned for the future.

* * *

Neurology/Neurosurgery

Perhaps the most optimistic finding to date has been the large size of many of the brain wounds in our group and the amazing ability of many of these young men to compensate for their injuries. CT scanning now shows us that 80% had injuries involving multiple lobes of the brain, and in 33% the injury was bilateral (the injuries were thus much larger than had been previously estimated from surgical reports and skull x-rays alone). To the casual observer, almost two-thirds of these patients might appear to be functioning normally. Nevertheless, careful examination almost invariably reveals some neurological or neurobehavioral functional deficit, and a review of family and community adjustment often reflects these abnormalities. Unrecognized cognitive and especially memory deficits often resulted in a failure to seek help or veterans' benefits; and many

patients with severe wounds had been returned to duty and eventually received nonmedical discharges from the military services. Thirty-eight percent of our brain-injured patients received a recommendation for psychological intervention or therapy, although many had previously undergone such therapy before participation in the VHIS. Additionally, about 28% of the controls also received such recommendations while at the VHIS. Overall, recommendations for neurologic or psychological follow-up were made in 72% of the brain-injured patients and 52% of the controls; the brain-injured group received more recommendations per person than the controls. We expect that much of our analysis will allow us to provide a series of guidelines that will use the initial CT scan and examination to predict eventual outcome, provide such patients and their families some insight into difficulties that they may expect, and target specific therapies for them early in their convalescence.

* * *

A 15-year mortality study was done on the 1127 men with penetrating craniocerebral injuries in the registry who were alive 1 week postinjury. During this time, 90 deaths (8%) occurred. Most of the deaths occurred early in the first year after trauma and were secondary to the direct effects of brain injury or the sequelae of coma. Complications, particularly infections, were significant mortality factors. Coma was the best prognostic guideline. Posttraumatic epilepsy was not related to mortality except for the risks accompanying each ictus. The population now appears to be approaching the actuarial norm of their peers (Rish, et al, J Neurosurg).

* * *

Epilepsy has long been recognized as one of the most troubling sequelae of brain trauma. Fifty-three percent of VHIS patients had developed epilepsy by 15 years posttrauma; and 50% of those were still having seizures at that time. Their relative risk of developing seizures in the first year postinjury was 640 times greater than for the normal age-matched population; and at 15 years postinjury, it was still 25 times greater. Patients with focal neurologic signs or large lesions had increased risk of epilepsy, and site of the lesion may have been more important than size in determining occurrence. Family history of epilepsy or level of preinjury intelligence had no effect on seizure occurrence. Seizure frequency in the first year predicted future severity of seizures. Phenytoin therapy in the first year after injury did not prevent later seizures (Salazar, et al, Neurology).

In a study of the anatomic correlates of epilepsy, structures or combinations of them involved on CT were screened for their Spearman rank correlation with the occurrence of epilepsy, and the 26 most highly correlated ($p < .05$) were allowed to interact in a backward multiple logistic regression analysis using brain volume loss as a covariate. This resulted in a model which correctly predicted the occurrence of epilepsy 71% of the time and to which only five "structures" contributed significantly ($p < .04$). These were: Left temporal white matter (relative risk = 3.3:1), right vertex gray (2.7:1), left convexity gray (2.4:1), right frontal white (2:1), and right corona radiata (1.8:1), in that order. The incidence of epilepsy reached 92% in patients with lesions in certain combinations of these structures. While total brain volume loss alone was associated with epilepsy, it did not significantly add to the model in the presence of these structures (Salazar, et al, Neurology).

We have also studied the cognitive and behavioral correlates of posttraumatic epilepsy in these men. The performance of head-injured epileptics and nonepileptics and 76 matched, uninjured controls on 16 neuropsychological measures was compared in ANOVA with and without correction for total brain volume loss on CT and within hemisphere of involvement. Measures included IQ, Wisconsin Card Sorting, selective reminding, word recognition, Kimura Recurring Figures, visual retention, finger tapping, and continuous performance tests as well as the Beck Depression Inventory and a community adjustment score. As expected, epileptics (who generally had larger lesions) performed more poorly, were more depressed and more poorly adjusted than head-injured nonepileptics, and both did more poorly than uninjured controls. However, after correction for lesion size, there were no significant differences between epileptic and nonepileptic men on most of the measures studied. Exceptions were the performance IQ, selective reminding, and finger tapping tests. This suggests that epilepsy per se is not responsible for most of the cognitive and psychosocial deficits seen in our head-injured patients. The roles of seizure type and medications are being explored (Salazar, et al, Neurology).

We are currently developing a formula based on time postinjury and integrating the above findings in order to predict the risk of epilepsy even more accurately in a given patient and thus target those patients most likely to benefit from prophylactic anticonvulsants (Weiss, et al, Arch Neurol). Future studies will pursue the above findings in greater detail, particularly the anatomic and cognitive correlates. Much information has also been collected on medication usage and seizure symptomatology. We have found a positive correlation of epilepsy with retained metal fragments which was surprising and must be investigated further, particularly because of its

implications for surgical treatment. The relation of epilepsy to motor and language function, a more detailed analysis of the extensive genetic information available to us, and analysis of the relation of epilepsy to the extensive social adjustment data available also remain to be pursued.

* * *

Forty-seven percent of our patients were recorded as having a paralysis early after injury, and about half of those have now recovered. Analysis of the clinical and anatomic correlates of recovery from hemiparesis has resulted in a simple initial model that may allow us to predict which patients will recover. Clinical findings significantly ($p \leq .05$) associated with nonrecovery were sensory loss, organic mental disorder, abnormal EEG, partial simple seizures, and an initial extensor plantar response. Anatomic correlates included large total brain volume loss and involvement of the following anatomic structures on CT: sensory-motor cortex, supplementary motor area, posterior temporal cortex, temporal white matter, posterior limb of internal capsule and corona radiata, lentiform, thalamus, and caudate. These clinical and anatomic factors were then allowed to interact in a stepwise logistic regression model comparing unrecovered patients to those with delayed recovery (>1 month postinjury). Items significantly ($p < .05$) predicting recovery in this model were CT scan involvement of (1) vertex or medial sensory motor cortex and (2) central corona radiata and caudate body; (3) extensor plantar response, and (4) sensory loss, in that order. Probability of recovery was .05 for patients with all items present and .97 when all were absent. This model was 82% accurate (Smutok, et al, Neurology).

Most patients who are going to recover motor functions will do so within the first 6 months after injury (15% recover within 1 month), but a small percentage may not do so for several years; we expect to be able to identify such patients early after injury. Other analyses have shown that considerable ipsilateral as well as contralateral deficits in complex hand motor functions can be found in patients with lesions in the frontal and parieto-occipital lobes even in the absence of an overt hemiparesis. This is most pronounced in patients with right hemisphere brain injuries and in right-handed individuals. Preliminary analysis also shows that the relation of persistent hemiparesis to eventual successful community adjustment is not direct and that other factors, primarily cognitive status, may play a more important role than paralysis per se. Follow-up studies will clarify this relationship. Analyses of the pattern

of motor recovery, of the relation of paresis to language function, as well as of the relation of spasticity to lesion location are also planned.

* * *

A separate study now planned in collaboration with the Medical Neurology Branch of the NINCDS in Bethesda, MD, will involve reevaluation of selected hemiparetic patients using Positron Electron Tomography (PET) scanning and topographic EEG mapping to elucidate the mechanisms of recovery from hemiparesis as well as from other neurologic deficits.

* * *

Analysis of consciousness and traumatic amnesia in our men showed that only 15% had prolonged unconsciousness and 53% had no or momentary unconsciousness after injury, emphasizing the focal nature of these wounds. There was a clear dominance of the left (or language-dominant) hemisphere for the "wakefulness" or vigilance component of consciousness. The area of the posterior limb of the left internal capsule, the left basal forebrain, midbrain, and hypothalamus were most associated with unconsciousness. Left dominance is not seen for posttraumatic amnesia after elimination of the "wakefulness" variable, suggesting that the latter may be linked to the well-recognized role of the left hemisphere in certain verbal memory processes (Salazar, et al, Neurology).

This particular analysis illustrates another example of the functional asymmetry of the two halves of the brain and has also helped to sharpen the distinction between the two major aspects of the arousal component of consciousness: "wakefulness" (left hemisphere) and attention (right hemisphere). Future studies on this subject will pursue the anatomic correlates of unconsciousness in more detail as well as the relation of this "wakefulness" component of consciousness to attention and their respective relationships to the various forms of memory failure.

* * *

Another example of how a study of these patients can contribute to seemingly unrelated branches of neurology is the analysis of men with injuries to the basal forebrain, a poorly understood region of the brain which has been shown to degenerate early in patients with Alzheimer's presenile dementia. Some investigators have thus suggested that the neuronal loss in the

basal forebrain may lead to dementia. The neurologic and cognitive performance of 15 young veterans who suffered unilateral penetrating missile wounds to the basal forebrain was compared to that of patients without basal forebrain lesions and uninjured controls. While they did somewhat more poorly on tests of episodic memory, reasoning, and arithmetic and had more prolonged unconsciousness postinjury, their performance compared favorably with that of uninjured controls on tests of intelligence, attention, and language and was not consistent with that of a demented patient. These results suggest that the basal forebrain may be a component of limbic-hippocampal memory processing systems but is not responsible for the maintenance of cognitive processing in general. Future analysis will pursue basal forebrain function in a broader range of memory and attention tasks (Salazar, et al, Neurology).

* * *

We have also studied the relationship between EEG findings, clinical and radiological features of the first 300 VHIS subjects. EEGs were performed on 16 and 18 channel Grass equipment using international 10-20 system. Fifty age-matched Vietnam veterans were used as controls. Electroencephalogram was abnormal in 48% of the patients. Epileptiform findings (EF = spikes or spikes wave) were found in 15% of the records and focal slowing (FS) in 38%. Eighty percent of the patients with EF had one or more seizures after head injury compared with 64% for FS and 41% for normal EEGs. EF were seen in 16% of patients who had their initial seizure during the first year following head injury but in only 7% of those with onset after 5 years. EEG was normal in 31% of the former and 71% of the latter group. No correlation was found between EF and family history of epilepsy, seizure frequency in first year after injury, or seizure persistence. Both EF and FS correlated significantly with hemiparesis ($p = 0,0001$), aphasia ($p = 0,00074$), and CT scan evidence of deep cerebral injury ($p = 0,0004$) (Jabbari, et al, Neurology).

Another analysis studied the relationship between visual evoked potentials (VEP), perimetry, clinical, and CT findings of the first 150 patients in our study. Full field (FF) and half field (HF) responses were obtained by a TV delivering checkerboard pattern reversal stimuli at a rate of 2.1/second. Responses were recorded on four medial and lateral occipital electrodes simultaneously, placed 5 and 10cm from the midline. Visual fields were obtained by Goldmann perimetry and CT scans by a GE 8800 scanner. Fifty age-matched Vietnam veterans served as controls. Fourteen patients (9%) showed a mono-occular delay of VEP on the side of head injury. Seven of these patients had no visual complaints, suggesting that VEP detected a subclinical traumatic

macular or optic nerve dysfunction. HF stimulation and perimetry produced concordant data in 88% of the patients. When abnormal, both tests correlated highly with a parieto-occipital site of injury. In six patients, abnormality of HF-VEP pointed correctly to the side of head injury but perimetry was normal; while in a few patients, perimetry showed small hemianopic field defects and HF-VEP missed them. This data indicates that HF-VEP is a sensitive measure of optic radiation dysfunction in penetrating head injury. Information derived from HF-VEP and perimetry complement each other in retrochiasmatic brain lesions.

* * *

In the area of audiology testing of the VHIS subjects, central auditory tests were administered to 250 individuals who had sustained penetrating head injury. Each subject also received computerized tomography for which a normal or abnormal rating was assigned for eight different regions of the temporal lobe. The location and degree of temporal lobe injury was compared to dichotic speech test results in an effort to establish auditory correlates of physical damage. Results indicate that speech test scores are significantly affected by injury site. In addition, three dichotic speech tests (SSW, dichotic digits, and dichotic CVs) were administered to 300 individuals with brain injury in various locations. The sensitivity of each test was studied relative to the percent of normal/abnormal scores for specific injury groups. A high rate of false negative and false positive results was present for all measures. The three dichotic tests did not vary substantially in their ability to detect right or left temporal lobe damage.

Time compressed speech has been reported to be a useful test in the identification and differentiation of central auditory deficits; therefore another analysis performed on 250 individuals with predominately discrete brain injury reports the results for 60% compressed NU #6 word lists. Absolute compressed speech scores, different scores (noncompressed minus compressed), and difference score ratios were obtained for each subject. Results indicate that this speech test lacks sensitivity to reliably identify or separate brain injury for the VHIS population (Sedge, Mueller, et al).

* * *

The principal neurosurgical question addressed to date has been the controversial issue of the significance of retained bone. The experience of previous wars had suggested that bone fragments retained in a brain wound served as a nidus for infection and increased morbidity and mortality. It thus became standard operating procedure in Vietnam to remove such fragments surgically, even if this called for repeated brain operations in otherwise healthy, convalescing patients. Over 10% of our patients thus underwent repeat surgery for this purpose, some of them multiple times. Retrospective analysis of CT scans now shows that almost 75% of those who had such surgery still have retained bone fragments, as do over 20% of the VHIS population. A detailed review of the medical records of each of these men shows that retained bone per se has no significant effect on mortality, morbidity (including infection rate), or sequelae of brain injury in this population. This strongly suggests that repeat operations for retained bone, in the absence of complications, are not warranted and may be detrimental. A similar analysis of the possible role of retained metal is planned (Myers, et al, in preparation).

Thanks in large part to helicopter evacuation and the deployment of neurosurgeons close to the battlefront, a wounded soldier in Vietnam usually received prompter and better medical care than was available anywhere in the world at that time for such wounds. Most men had definitive neurosurgery within 6 hours of injury, but a preliminary analysis of complication rates by delay in provision of initial surgery shows that mortality and morbidity begin to rise significantly only with delays longer than 24-48 hours. Combined with data on early hospital mortality, this type of information may be important for medical logistical and evacuation policy planning in future conflicts (in which such prompt care may not be possible). An Army-sponsored project designed to develop new medical adjunctive treatments which may help minimize or delay the need for prompt definitive neurosurgery and minimize tissue loss is currently being planned, using VHIS findings and evaluation techniques as a springboard.

Other neurosurgical questions which remain to be addressed in the data include the relation of ventricular enlargement to intraventricular wounds, to clinical and cognitive deficits, and to eventual community adjustment, and the relation of surgical complications such as infection to wound type, fragment type and size, surgical procedure, spinal fluid leaks, and eventual outcome.

* * *

Other planned analyses of the neurological data include a study of headaches; the anatomic, clinical, and evoked potential correlates of the various types of sensory loss and their relation to community adjustment; analysis of specific visual field deficits; and evaluation of the effect of various therapies including anticonvulsants, speech therapy, physical therapy, and psychiatric intervention.

* * *

Neurobehavior

Initial analyses of the speech and language data have included a study of recovery from Broca's aphasia and a study of speech discrimination deficits. The first was designed to determine which language faculties are retained in the chronic form of expressive aphasia and what characteristics of brain lesions differentiated between patients who recovered and did not recover from expressive aphasia within 15 years following penetrating head injury. Two groups of men who sustained penetrating head injuries and had an expressive aphasia during the first 6 months following injury were examined 15 years later: one group had a chronic expressive nonfluent aphasia and the other had recovered and was without symptoms of aphasia. On a comprehensive battery of speech and language tests, the chronic expressive aphasics demonstrated specific deficits in syntactic processing in all language modalities, while being within normal range in other language faculties. The recovered group demonstrated syntactic deficits only in written expressive syntax. The CT lesions of the two groups differed in the extent of left hemisphere lesion volume and the degree of posterior and deep lesion extension. Broca's area was equally involved in both groups but was not involved in all patients in either group. All the nonrecovered group had posterior extension of their lesion to involve Wernicke's area with some involvement of the underlying white matter and basal ganglia in the left hemisphere (Ludlow, et al, in press, Brain).

Speech discrimination and identification tasks assessing voicing and place distinctions were given to 16 unilaterally brain-injured subjects free of aphasic or dysarthric symptoms 12 to 15 years postinjury. Seven subjects did not demonstrate any difficulty with these speech tasks, while five left and four right brain-injured subjects showed moderate difficulties. These difficulties were more pronounced on the discrimination than on the identification tasks. Analysis of CT scans demonstrated that the lesion locations most clearly associated with the speech discrimination deficits were upper levels of the white matter subjacent to cortical regions in either hemisphere (Yeni Komshian, et al, in press). Other analyses under way will study recovery from Wernicke's aphasia and patients with dysprosody.

Neuropsychology: The Neuropsychology Section (Drs. Jordan Grafman and Herbert Weingartner) of the VHIS was developed to broadly address certain critical issues regarding brain-behavior relationships, the conceptual validity of specific cognitive theories, and the persistence of cognitive deficits and their effect upon the clinical course of a patient. We began by assessing the impact of education, preinjury intelligence, brain volume loss, and lesion location upon postinjury intelligence level (Grafman, et al, Science, in press). We found that left hemisphere lesions and lower preinjury education had a profound effect upon postinjury intelligence performance. This finding was not surprising given the linguistic processing demands of the Armed Forces Qualification Test (AFQT). In addition, we discovered that the more global a cognitive process measured, the greater the effect of brain loss volume--specific cognitive processes were affected relatively more by lesion location. This illustrates a methodological approach that will continue to guide our research effort: distinguishing between effects on global vs. specific cognitive/mood processes by considering both anatomical and behavioral variables.

An example of a specific cognitive process might be the ability to discriminate and recognize faces. Our analysis indicates that both hemispheres of the brain contribute to this process, with the left hemisphere storing face knowledge information and the right hemisphere storing procedures that allow for rapid face discrimination and form memory. Face recognition that requires transformation of features (e.g., the person has to rely on specific face features for recognition) seems to require the integrity of the frontal lobe (Grafman, et al, submitted for publication). A second example of a specific cognitive process involves the semantic encoding of recently presented verbal information. We have tested an individual who presented with a restricted deficit in this process in contrast to superior skills in all other cognitive areas. We argue that his critical lesion is to the columns of the fornix (Grafman, et al, submitted for publication).

We have taken a parallel course in examining the mood presentation of our patients. A particularly interesting area of investigation is the effects of frontal lobe lesions upon the maintenance of control of anxiety, fear, and hostility. We have demonstrated (Grafman, et al, submitted for publication) the rather profound and persistent effects of orbitofrontal lesions upon the modulation of feelings of anxiety, dorsofrontal lesions upon feelings of sluggishness, and the acute effects of frontal lobe lesions in general upon control of anger and hostility. Patients with left dorsofrontal and right orbitofrontal lesions were most disinhibited, edgy, angry, and depressed. Studies currently in progress investigate single cases with limited

orbitofrontal lesions, Beck Depression Inventory group profiles, MMPI group profiles, and factor analysis of the Bear-Fedio Trait Scales. Our purpose is to develop a rudimentary model of mood state representation and to discover how mood state interacts with cognitive processes.

A third line of inquiry, just beginning, looks at injury variables (e.g., epilepsy, consciousness, and motor functions) and how they will affect subsequent cognitive functioning.

A fourth analysis scheme is clinical in nature and tries to address the functional significance of penetrating brain wounds. An initial analysis (Dresser, Grafman, et al, submitted for publication) indicated that while more severely disabled subjects made greater use of available benefits, a substantial number of 100% disabled (V.A. rated) men did not use benefits which would appear to have aided their functional adjustment. We will be pursuing this issue further.

Thus, we have begun to exploit the VHIS neuropsychological data in several separate areas: cognition, mood, injury characteristics, and functional/clinical outcome utilizing lesion location, brain loss volume, and preinjury intelligence as covariates. We believe that the initial studies in each area will not only contribute to the scientific and clinical literature, but provide the basis for continuing analysis in the future. This continuing analysis is necessary in order to refine the models of brain/behavior we have only barely begun to construct. Some of the evolving research directions are itemized as follows:

1) We plan to analyze the interdependency of seemingly dissimilar cognitive processes via clustering, factor analysis, and correlational techniques. The purpose will be to assess how brain injury affects this interdependency (e.g., reasoning could be viewed as the computational skill that underlies task interdependency).

2) We plan to develop in more detail a model of mood regulation identifying key cortical structures and cognitive processes, focusing on the idea that mood sensation and cognitive explanation/regulation of mood are parallel processes that can be either fractionated or made more interdependent by brain injury.

3) We plan to more fully present a picture not only of the psychosocial course of the patient, but the impact of the patient's head injury upon family members.

4) We have planned and are currently conducting a series of studies on memory processes in order to develop a functional model of memory and to consider how brain lesions affect memory in a

young adult population. We are currently analyzing three aspects of memory processing--the distinction between long-short, episodic-semantic, and automatic-effortful processes.

5) We plan to investigate whether so-called right hemisphere brain functions are more diffusely represented compared to left hemisphere brain functions.

6) We plan to develop a clinical management model for the long-term care and educational development of the head-injured young adult.

7) We plan to investigate and better explain the computational role of the frontal lobes of the brain.

8) We plan to identify both the perceptual and language functions of the right hemisphere.

Thus, in addition to collaborations with individual investigators, we have a 3-4 year plan of further neuropsychological data exploitation. We expect this research will continue to develop neuropsychological theory and help us to offer practical suggestions for the management of the head-injured young adult.

* * *

SUMMARY

In summary, the VHIS data base represents an invaluable asset on computer tape and microfiche that will continue to provide room for analysis for years to come. While many of the questions posed in the original protocol have already been answered, new and often more exciting questions have arisen and will continue to arise as investigators explore the data. The VHIS evaluation has also served to identify subsets of patients with specific types of wounds or deficits who can be invited to return for more detailed experimental testing that concentrates on their specific disabilities or on hypothesized functions of the brain areas involved in their injuries. Many of the questions have immediate practical implications for prediction of outcome, for therapy, and for determination of disability status. However, perhaps the most valuable aspect of the study will be the long-term benefits resulting from a better scientific understanding of brain function and its localization.

VIETNAM HEAD INJURY STUDY

Manuscripts, Abstracts, Presentations
(As of 9 July 1985)

MANUSCRIPTS

Published:

Caveness WF, Meirowsky AM, Rish BL, Mohr JP, Kistler JP, Dillon JD Weiss GH. The nature of posttraumatic epilepsy. J Neurosurg 50:545-553, 1979.

Rish BL, Dillon JD, Meirowsky AM, Caveness WF, Mohr JP, Kistler JP, Weiss, GH. Cranioplasty: A review of 1030 cases of penetrating head injury. Neurosurg 4(5):381-385, 1979.

Meirowsky AM, Caveness WF, Rish BL, Dillon JD, Mohr JP, Kistler JP, Weiss GH. Definitive care of cerebral missile injuries crossing the midline. Military Medicine, 145(4):246-250, April 1980.

Mohr JP, Weiss GH, Caveness WF, Dillon JD, Kistler JP, Meirowsky AM, Rish BL. Language and motor disorders after penetrating head injury in Vietnam. Neurol, 30(12):1273-1279, December 1980.

Rish BL, Dillon JD, Caveness WF, Mohr JP, Kistler JP, Weiss GH. Evolution of craniotomy as a debridement technique for penetrating craniocerebral injuries. J Neurosurg 53:772-775, 1980.

Rish BL, Caveness WF, Dillon JD, Kistler JP, Mohr JP, Weiss GH. Analysis of brain abscess after penetrating craniocerebral injuries in Vietnam. Neurosurg 9(5):535-541, 1981.

Meirowsky AM, Caveness WF, Dillon JD, Rish BL, Mohr JP, Kistler JP, Weiss. Cerebrospinal fluid fistulas complicating missile wounds of the brain. J Neurosurg 54:44-48, 1981.

Meirowsky AM. Secondary removal of retained bone fragments in missile wounds of the brain. J Neurosurg 57:617-621, 1982.

Weiss GH, Feeney DM, Caveness WF, Dillon JD, Kistler JP, Mohr JP, Rish BL. Prognostic factors for the occurrence of posttraumatic epilepsy. Arch Neurol 40:7-10, Jan 1983.

(Published manuscripts continued)

Rish BL, Dillon JD, Weiss GH. Mortality following penetrating craniocerebral injuries: An analysis of the deaths in the Vietnam Head Injury Registry population. J Neurosurg 59(5):775-780, 1983.

Sweeney JK & Smutok MA. Preliminary analysis of the functional and anatomical sequelae of penetrating head trauma. J Amer Phys Ther Assn 63(12):2018-2025, Dec 1983.

Ludlow CL. The brain bases for language functioning: New insights from penetrating head injuries. In: Georgetown University Round Table on Languages & Linguistics, 1982. Edited by Heidi Byrnes, Washington DC: Georgetown University Press, pp 203-223.

Ludlow CL. Identification and assessment of aphasic patients for language intervention. In: Contemporary Issues in Language Intervention. Edited by J. Miller, D.E. Yoder, & R. Schiefelbusch. ASHA Reports No. 12, Rockville, MD: American Speech-Language-Hearing Assn, 1983, pp 75-91.

Beck WG, Mueller HG, Sedge RK. A measure of test-retest reliability of the SSW. SSW Reports 7(2):8-10, May 1985.

In Press:

Salazar AM, Jabbari B, Vance SC, Grafman J, Amin D, Dillon JD. Epilepsy after penetrating head injury I: Clinical correlates. Neurology.

Salazar AM, Grafman J, Vance SC, Weingartner H, Dillon JD, Ludlow CL. Consciousness and amnesia following penetrating head injury. Neurology.

Grafman J, Salazar AM, Weingartner H, Vance SC, Ludlow CL. Isolated impairment of memory following a penetrating lesion of the fornix. Archives of Neurology.

Grafman J, Salazar AM, Weingartner H, Vance SC, Amin D. The relationship of volume loss and lesion location to cognitive deficit. Journal of Neuroscience.

(In press manuscripts continued)

Grafman J, Smutok M, Sweeney J, Vance SC, Salazar AM, Weingartner H. The effects of left hand preference on postinjury measures of distal motor ability. Perceptual and Motor Skills

Ludlow C, Rosenberg J, Fair C, Buck D, Schlesselman S, Salazar AM. Brain lesions associated with chronic expressive syntactic aphasia following penetrating head injury. Brain.

Mueller G, Sedge RK, Salazar AM. Central auditory nervous system dysfunction: Preliminary findings of the Vietnam Head Injury Study. In: Miner ME, Wagner KA (eds), Neural Trauma: Treatment, Monitoring, and Rehabilitation Issues. Butterworth Publishers, Stoneham, MA, 1985.

Grafman J, Salazar AM. Methodological considerations relevant to the comparison of recovery from penetrating and closed head injuries. In: Levin H, Eisenberg H, Grafman J (Eds.), Neurobehavioral Recovery from Head Injury, New York: Oxford University Press, 1985.

Levin H, Eisenberg H, Grafman J. Neurobehavioral Recovery from Head Injury, Oxford University Press, 1985.

Submitted:

Salazar AM, Jabbari B, Grafman J, Amin D, Smutok M. Anatomic and clinical correlates of posttraumatic epilepsy.

Salazar AM, Grafman J, Schlesselman S, Vance SC, Carpenter M, Pevsner P, Ludlow CL, Weingartner H, Mohr JP. Penetrating war injuries to the basal forebrain: Neurologic and cognitive correlates.

Grafman J, Weingartner H, Salazar AM, Ludlow C, Amin D, Dillon JD. Intellectual function following penetrating head injury in Vietnam veterans.

Grafman J, Salazar AM, Weingartner H. Face memory and discrimination: A preliminary analysis of the persistent effects of penetrating brain wounds.

Grafman J, Vance SC, Weingartner H, Salazar AM, Amin D. The effects of lateralized frontal lesions upon mood regulation.

(Submitted manuscripts continued)

Grafman J, Vance SC, Weingartner H, Salazar AM. Specific effects of orbitofrontal brain wounds upon regulation of mood.

Grafman J, Smutok M, Salazar AM, Vance SC, Sweeney J, Amin D. The persistent effects of penetrating brain wounds upon simple distal motor skills in right-handed men.

Grafman J, Salazar AM, Vance SC, Weingartner H, Ludlow C, Amin D. Immediate memory for story discourse in Vietnam veterans with penetrating brain wounds.

Weingartner H, Grafman J, Salazar AM. Relationship between depressed mood and cognition in patients with lateralized brain lesions.

Dresser A, Grafman J, Salazar AM, Hoyt M, Smutok M, Brown H. A preliminary analysis of benefits: The Vietnam Head Injury Study.

Yeni-Komshian G, Ludlow CL, Rosenberg J, Fair C, Salazar AM. Lesion locations associated with speech deficits following penetrating head injury.

West G, Grafman J, Christodoulou C. The relationship between laterality of lesion and MMPI scale scores.

Ludlow CL, Rosenberg J, Salazar AM, Grafman J, Smutok M, Buck D, Dillon JD. Persistent speech dysprosody following penetrating head injury.

Boehm TM, Salazar AM. Hypothalamic-pituitary function in severely head-injured Vietnam veterans.

ABSTRACTS

Salazar AM, Dillon JD, Mohr JP, Meirowsky AM, Weiss G, Rish G. Penetrating head injury in the Vietnam War: Relation of initial neurological status and anatomic loss to long-term outcome. Neurol 32(2) A127, 1982.

Salazar A, Jabbari B, Dillon D, Grafman J. Seizures following penetrating head injury in the Vietnam War, a preliminary report. Neurol 33(2):215, 1983.

(Abstracts continued)

Salazar AM, Grafman J, Mohr J, Dillon JD, Pevsner P. Penetrating war injuries to the basal nucleus of Meynert: Anatomic, neurologic, and cognitive correlates. Neurol 33(2):104, 1983.

Salazar AM, Jabbari B, Grafman J, Amin D, Smutok MA. Anatomic and clinical correlates of posttraumatic epilepsy. XV Epilepsy Intl Symposium, Washington, DC, September 1983.

Grafman J, Salazar AM, Weingartner H, Dillon J. Effects of temporal lobe lesions due to penetrating missile wounds upon selected cognitive measures. Neurol 33(Suppl 2):218, 1983.

Grafman J, Salazar AM, Weingartner H. Unfamiliar face recognition and discrimination. A preliminary analysis of the persistent effects of penetrating brain wounds. Intl Neuropsychol Soc, Lisbon, 1983.

Grafman J. Cognitive sequelae and functional outcome in penetrating missile wound injury. International Conference on the Management of Traumatic Brain Injury, London, England, July 1983.

Jabbari B, Vengrow M, Salazar AM, Grafman J, Dillon JD, Gunderson CH. Clinical significance of EEG findings in penetrating head injury. A preliminary report on 300 Vietnam veterans. Neurol 33(Suppl 2):188, 1983.

Jabbari B, Salazar AM, Smutok M, Amin D, Gunderson CH. Clinical significance of somatosensory evoked potential findings in penetrating head injury. Am EEG Soc, 1983.

Myers PW, Salazar AM, Dillon JD, Meirowsky AM, Schlesselman S. The significance of retained intracranial bone fragments. Congress Neurol Surgeons, September 1983; Neurosurg, 1983.

Mueller HG, Sedge RK, Salazar AM. SSW results in head injury: Preliminary report on 200 cases. Am Speech-language Hearing Assoc, Toronto, 1983.

Mueller GH, Sedge RK, Salazar AM. Head injury and auditory processing: Sensitivity of three dichotic tests. Am Speech-Language Hearing Assoc, 1983.

(Abstracts continued)

Mueller GH, Sedge RK, Salazar AM. Head injury and auditory processing: Functional and physical correlates. Am Speech-Language Hearing Assoc, 1983.

Vance SC, Grafman J, Jabbari B, Salazar AM. Posttraumatic language-induced epilepsy. Am Acad Neurol, April 1984.

Vance SC, Grafman J, Weingartner H, Salazar AM. Specific effects of orbitofrontal brain wounds upon regulation of mood. Am Acad Neurol, April 1984.

Salazar AM, Grafman J, Vance SC, Schlesselman S, Weingartner H. Unconsciousness and amnesia following penetrating head injury: Neurologic, cognitive, and anatomic correlates. Neurol 34(1):233, 1984.

Salazar AM, Grafman J, Jabbari B, Vance SC, Amin D. Epilepsy and cognitive loss after penetrating head injury. XV Epilepsy Intl Symposium, Hamburg, Germany, Sept 1985.

Salazar AM, Amin D, Vance SC, Schlesselman S, Buck D, Grafman J. Seizures after penetrating head injury: Effects of lesion location - Anatomic correlates. XV Epilepsy Intl Symposium, Hamburg, Germany, Sept 1985.

Salazar AM, Grafman J, Jabbari B, Vance SC, Amin D. Epilepsy after penetrating head injury: Cognitive correlates. Am Acad Neurol, April 1985.

Salazar AM, Amin D, Vance SC, Schlesselman S, Buck D, Grafman J. Epilepsy after penetrating head injury: Anatomic correlates. Am Acad Neurol, April 1985.

Smutok MA, Vance SC, Salazar AM, Foulkes M, Grafman J. Neurologic and anatomic correlates of recovery from hemiparesis following penetrating head injury. Am Acad Neurol, April 1985.

Grafman J, Ludlow CL, Weingartner H, Salazar AM. The persistent effects of penetrating brain injury upon the accessability of "semantic" versus "episodic" information. Intl Neuropsychol Soc, 1985.

Grafman J, Vance SC, Weingartner H, Salazar AM, Amin D. The effects of lateralized frontal lobe lesions upon mood regulation. Intl Neuropsychol Soc, 1985.

PRESENTATIONS

Ludlow CL. The brain bases for language functioning: New insights from penetrating head injuries. 33rd Annual Georgetown University Round Table on Language and Linguistics, Washington, DC, March 1982.

Ludlow CL. Identification and assessment of aphasic patients for language intervention. National Conference on Language Intervention, Omaha, NE, April 1982.

Ludlow CL, Rosenberg J, Dillon D, Buck D. Head injuries associated with persistent speech dysprosody. Speech Motor Control Conference, Madison, WI, April 1982 and Academy of aphasia, New York, October 1982.

Grafman J, Smutok M, Sweeney J, Weingartner H. Hemispheric representation of simple distal motor processes. International Neuropsychological Society Meeting, Deauville, France, 1982.

Grafman J, Weingartner H. Text processing in brain-lesioned patients. International Neuropsychological Society Meeting, Deauville, France, 1982.

Salazar AM. Vietnam Head Injury Study. NATO Conference on Head Injury, London, England, June 1983.

Salazar AM. Vietnam Head Injury Study. Grand Rounds, Department of Neurosurgery, Glascow University Hospital, Scotland, June 1983.

Salazar AM. Vietnam Head Injury Study. Presentation to Department of Neurology, Radcliff Infirmary, Oxford, England, June 1983.

Grafman J. Neuropsychological outcome of penetrating head injury: A preliminary analysis. Fourth Annual Traumatic Head Injury Conference: Braintree Hospital, Braintree, MA. October 1983.

Grafman J, Weingartner H, Dillon JD. Intellectual changes following penetrating missile wounds in Vietnam veterans. International Neuropsychology Society Meeting, Mexico City, 1983.

(Presentations continued)

Grafman J. Recovery of function following brain injuries: A perspective from penetrating head wounds. International Neuropsychology Society Meeting, Mexico City, 1983.

Yeni-Komshian G, Ludlow C, Rosenberg J, Fair C, Salazar A, Buck D. Speech discrimination deficits associated with penetrating head injury. Academy of Aphasia, Minneapolis, October 1983.

Smutok MA. Motor and functional recovery following focal brain trauma. Fifth Annual Traumatic Head Injury Conference: Braintree Hospital, Braintree, MA, October 1984.

Grafman J, Weingartner H, Salazar AM, Vance SC, Amin D. Frontal lobe lesions: Persistent effects upon cognition and behavior. International Neuropsychology Society Meeting, Houston, TX, February 1984.

Grafman J, Weingartner H, Salazar AM, Schlesselman S. Patterns of visual-spatial impairment: Evidence from penetrating brain wounds. International Psychology Society Meeting, Aachen, West Germany, June 1984.

Salazar AM. Vietnam Head Injury Study. Grand Rounds, Dent Neurologic Institute, Buffalo, NY, January 1985.

Smutok MA. Anatomic and clinical correlates of the recovery of motor function following focal brain injury. American Physical Therapy Association 1985 Combined Sections Meeting, Orlando, FL, February 1985.

Salazar AM. Vietnam Head Injury Study. Presentation to Washington Clinical-Pathological Society, April 1985.

Smutok MA, Grafman JH, Sweeney JK, Salazar AM. Ipsilateral effects of hemispheric lesions on upper extremity functions in hemiplegia. American Physical Therapy Association 61st Annual Conference, New Orleans, LA, June 1985.

Smutok MA. Outcome following penetrating head injury. Association of Military Surgeons of the USA (AMSUS), Anaheim, CA, November 1985.

6. Brain Mapping Factory

PARADIGM 2

Brain Mapping Factory

Characteristics of a Model Factory
(Semiconductor Wafer Fabrication)
(– L.G. Bailey/Veeco)

1. 5000 wafer starts per week, 5 day operation
2. 125 mm wafers, 4.8 inch usable diameter
3. 24 day cycle time
4. $36M equipment investment
   $750K depreciation/month
5. $15M plant investment
   $125K depreciation/month
6. Finished wafer to probe cost of $200

Flexible Material Handling Automation in Wafer Fabrication

James G. Harper and Louis G. Bailey
Veeco Integrated Automation, Inc., Dallas, Texas

The relationship between the need for increased integration of electronic functions, chip cost, and the value of automation is developed. High yield is the requirement for economic production of integrated circuits and the factors which contribute to low yield are delineated. Particulate contamination is a prime cause of yield loss and a significant portion of this contamination is contributed by people. Automation can decrease the number of people and particles and can improve yield as well as productivity, cycle time, and also can reduce inventory. Due to the continuing change in wafer fabrication factories, automation which is flexible in nature is the appropriate choice for material handling. It is shown that a flexible material handling system is cost effective.

SOME OF THE CHANGES which have occurred in wafer fabrication in the last few years relate to driving forces which lead to the need for flexible material handling automation. Although it is difficult to find positives in the 1980–1982 time frame for semiconductor manufacturers, or for semiconductor equipment manufacturers, some directions developed which can be viewed as positive. Among these are two which have direct relevance to material handling automation: the increased emphasis on manufacturing, and the increased development of standards, particularly the Semiconductor Equipment Communication Standards, SECS I and II. Both of these areas of attention play a role in generating an approach to material handling automation equipment in wafer fabrication factories.

The most prevalent characteristic in the thirty-odd year existence of the semiconductor industry has been change. That change manifests itself in many ways, but much attention always has been directed toward lowering the cost for a given electronic function [1]. In the last twenty years, the effort to lower cost has consisted primarily of two interrelated techniques: increased integration and ever-decreasing minimum feature size. This trend is delineated clearly by the overview presented in Table I.

Increased integration without feature size decrease is a very limited concept, due to the fact that yield decreases as chip area increases. The chip size is thus limited by economic considerations. By decreasing minimum feature size however, chip size has increased slowly over the years while the degree of integration has increased at a rate of about 4 × per four years. Minimum feature size as a function of time is shown in Fig. 1 [2]. The lower line represents the leading edge of state-of-the-art products, while the upper line represents an estimate of the position of peak capacity reached by the semiconductor industry. The knee of the curve occurs where lithography technology changes from visible projection systems to steppers or shorter wavelength sources. This subject has been discussed in the literature and will not be discussed here.

The importance of the two curves lies in showing the transition from the ability to produce small quantities of a new product to the ability to mass-produce that product at low cost. Another way to state the difference is that the upper

Table I—Integration Impact and Feature Size History

	Integration Step	Feature Size (μ)	Number of Components	System Cost ($)	Improvement Ratio
1	Discrete Components	25 0	30,000	10,000	Base
2	Integrated Circuits - SSI	10 0	1,000	1,000	10
3	Integrated Functions - MSI	7 0	200	500	20
4	Microprocessor - LSI	5 0	10	200	50
5	Microcomputer - VLSI	3 0	5	50	200
6	Single-Chip Microcomputer	1 0	1	20	500

curve requires high yield. Yield is the element which allows the upper curve to represent the peak capacity feature size. Note that in the future, the minimum feature size of peak capacity will grow closer to that of state-of-art products. The implication is clear: high yield must occur at smaller minimum feature size. This has been and will continue to be the reality of the semiconductor business.

Yield is the most critical parameter in the successful operation of a wafer fabrication operation. Nonetheless, several other factors play key roles in the cost of products produced by that operation. Among these are:

1. Labor productivity
2. Inventory control
3. Equipment utilization
4. Management philosophy.

These more traditional factors may be considered in the conventional method, i.e., payback. However, they can be considered only after all yield-related actions are implemented. The challenge remains achievement of high yield with ever-decreasing minimum feature size.

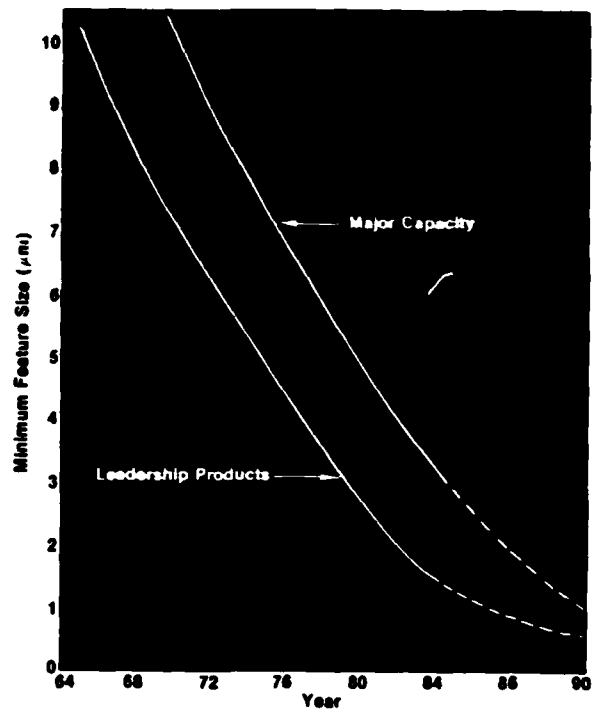


Fig. 1—Feature size trends.

Wafer Fabrication Design Criteria

If yield is the driving force, then the wafer fabrication factory must be designed in a manner which improves the opportunity to achieve high yield. To accomplish this goal, the causes of poor yield must be understood and systematically eliminated. Many reasons have been advanced for the production of scrap; among these several are worthy of note. They are:

1. Poor design
2. Poor relationship between design and test
3. "Bad Masks" (mask defects)
4. Inadequate processes (undercutting)
5. Physical or chemical impurities
6. Particles.

Some of the items on this list are not related to the quality of the wafer fabrication facility or the processes that are run in the facility. However, the last item, particles, is clearly related to the facility and the way in which it is designed and operated.

If a wafer fabrication facility is to achieve the goal of high yield, it must be designed and operated in a manner which reduces the number of particles on the wafers it processes [3]. The control of particles in the wafer environment is a major design criterion for fabrication facilities. Other design criteria essential to the design are:

1. The wafer input capacity desired
2. The desired processes
3. The equipment selected to run these processes
4. The desired work flow patterns
5. The methods of providing input chemicals and waste disposal
6. The structure of information flow
7. The control of personnel.

While each of these plays an important role in the ultimate configuration of a facility, they do not constitute the principal thrust of this discussion. Many of the above are decided on the basis of history and/or the preferences of the designers. However, the issue of particle control is an ingredient in all designs, no matter what the personal choices are in other areas.

Particle Sources

Particles come from many sources [4], but for the purposes of this discussion they may be categorized into five general areas:

1. The air in the facility
2. Input chemicals
3. The process itself
4. The process equipment
5. The people.

In general, a particle which is ⅓ to ¼ the minimum feature size will be a killing "defect." Therefore, a product with a minimum feature size of 1 micron should be protected from particles which have a nominal size of 0.3 microns [5].

The standard method of controlling airborne particles is through the use of clean rooms in which the air input is filtered. Usually the input air enters through filters in the ceiling at a nominal velocity of 100 ft/min and is exhausted through the floor or through the wall near the floor. The standard method of rating the quality of the air is related to the number of particles of a given size found in a cubic foot of air. This number is the class number. Thus, if 100 particles larger than 0.5 microns are measured in a cubic foot of air, the air is said to be class 100. Unfortunately this is the lower limit of the defined standard.

Several suppliers produce filters which filter very effectively above 0.3 microns. This produces excellent filtration at 0.5 microns and the resultant statements that a given air supply is class 10 or class 1. These classes are not defined, and more important, they usually do not specify particle size.

Current technology permits adequate filtration for products with feature size in the 1 micron range. However, this is true only if the time of exposure to filtered air is limited. One particle per cubic foot in a 100 ft/min air stream implies that 6000 particles per square foot per hour pass through a filter. A 150 mm wafer occupies approximately 19% of a square foot, and, if left flat, will intercept over 1000 particles in an hour of exposure. Of course, many other factors alter this conclusion. Slices are not left flat, but are placed vertically in cassettes; sticking coefficients and static charge attraction alter the number that stay, and cleaning prior to processing alters the number which cause yeild loss. In any case, the period of time during which wafers are exposed to moving air should be controlled if particle control is a key requirement [6].

In addition, to prove that the world is perverse, the boxes that store cassettes to remove wafers from the air stream cause problems due to outgassing and to difficulty in cleaning. This is an area which requires improvement in design and in materials selection, and will be discussed below.

Liquids and gases which provide an environment for the wafer should be of the same quality as that of the ambient air [7]. Filters which stop particles larger than 0.2 microns are available and should be placed in-line as point-of-use filters. Incoming quality control checks should also verify particulate levels. Piping should be pre-cleaned, as straight as possible, and without dead ends. Again, this is an area where attention to detail can provide adequate freedom from particulate contamination.

Processes by their very nature can produce particles which are detrimental. Examples of this type are plasma depositions or etching. These reactions form polymer-like materials which can coalesce into particles. The process engineer carefully controls process parameters to minimize unwanted reaction products in the area of the wafer. The role of the process engineer is critical to improvement. His role is also very interesting when material handling automation is to be considered, as it will be shortly.

In certain cases the interaction between process and equipment produces particles. Positive photoresist is brittle, and improper handling produces particles. The potential solutions are: (1) change the fracture characteristics of the photoresist or; (2) improve the material handling. Such particle-related problems require close interaction between the process engineer and the equipment engineer, to obtain rapid solutions.

The equipment must not add particles to the wafer. This cannot be stated too firmly. A shower of particles during a process reaction adds a discontinuity at the worst time. Equipment can be cleaned. Methods of verification are available to both the builder and user of equipment. If high yield is desired, then equipment which adds particles to the wafer must not be tolerated.

Unfortunately, the world of semiconductor manufacturing is inhabited by people. From the standpoint of small particle generation, even the clean people are dirty [8]. American Air Filter documents that a sitting person with light hand, head, and forearm movements will emit, in one minute, more than 500,000 particles larger than 0.3 microns. It is no wonder that clean rooms with excellent air near the ceiling have high particle counts at the wafer level where people work. Clean-room garments, properly maintained, help [9]. They reduce human particle emission by a factor of five, to the level of 100,000 particles per minute. Clearly the management of people-interaction with wafers is a key yield requirement in the design of a wafer fabrication factory. It is, in fact, one of the prime reasons for automation in wafer fabrication.

Significant effort has been directed toward process machines which operate without human intervention. While this effort has been directed primarily toward process reproducibility, it has had the definite benefit of removing the human from the wafer environment. This is a positive if the equipment is clean.

Estimates of the distribution of particle sources vary; however, experience indicates that the following average ranges are reasonable, if good design and operational technique are employed:

1. Air 5-10 (%)
2. Gases, liquids 5-10 (%)
3. Process 20-30 (%)
4. Equipment 20-30 (%)
5. People 30-40 (%)

The automated process machine limits the interaction of operators to loading and unloading. Veeco is pursuing development of automated loading and unloading equipment to eliminate this human/wafer interaction. This will allow the processing of wafers to be accomplished without direct human contact with a wafer or a wafer cassette.

The Force of Change

At the beginning of this discussion it was stated that the industry's historical trend has been change directed toward higher levels of integration and achieved by decreasing minimum feature size. This trend will continue, and it is this trend that dictates the shape of wafer fabrication facilities in the next several years. Process equipment that produced the 16K DRAM was inadequate for producing the 64K DRAM. New equipment is being used for the 256K DRAM. Equipment lifetime in a state-of-the-art wafer fabrication facility is similar to the rate of integration; a major change occurs every four years. If minimum feature size continues to decrease, and it will, then equipment will change to provide capability at the new feature size. New equipment is usually not completely perfected, and significant effort on the part of the user is required to achieve the desired process result. Many variations of process usually must be explored to achieve the desired result. The specific process changes according to the perceived capability of the equipment. Furthermore, the sequence of process steps changes. Process engineers are searching carefully for the right combination of variables to produce high yield. In fact, the only method process engineers have to improve processes is to change them. Combined with the fact that all process equipment is

not yet totally automated, this leads to the conclusion that the fundamental characteristics of wafer fabrication facilities will evolve slowly over the next several years.

The need for human capability in the factory remains. The skills, judgment, knowledge, and decision making which people bring to the factory all are required to produce leading-edge products. Equipment is not yet smart enough to eliminate the need for a helping human hand. Thus, even if the factory has automated process equipment and automated material handling, it will require people. What the automated equipment can do is to separate the people from direct contact with the wafers so that their presence does not cause yield loss.

The suggestion that wafer fabrication facilities will change very slowly over the next few years is disturbing to many in the semiconductor business. The cost of typical high-volume factories can be more than $50 million [10]; and the price of the equipment that is required to process sub-two micron products is rising, so this cost does not appear to be stablizing. Over the last fifteen years, several attempts have been made to alter fundamentally the clean-room approach to facilities. A method representing significant change and attempted by several groups employs process machines linked by minimum-size clean air tunnels, as exemplified by the IBM QTAT approach. These approaches have met with limited success. They have achieved the understanding that some process sequences are logically directly linked.

A good example of logically linked steps would be the photoresist coat-bake sequence. These sequences have been expanded to include scrub at the start and print-develop-bake-etch at the output. However, linked equipment suffers from the tyranny of uptime when too many steps are linked. The problem is that if one piece of equipment requires attention, then the output of several pieces of equipment degrades. In short, logically linked equipment chains are appropriate, but the nature of wafer fabrication is such that buffers and optional processing equipment are of value in optimizing output. Approaches which provide alternatives to clean rooms inhabited by people will continue to be investigated. They will catch on slowly because the majority of the cost of a factory lies in the cost of equipment, and capacity or capability is usually required quickly. This places the new facility concept in the role of a risky venture, which will cost a great deal of time (sales) and money if unsuccessful.

Upcoming Directions

We have discussed some issues to be considered in wafer fabrication design and operation. Let us briefly summarize and then proceed to a description of how automated material handling should be configured in a factory.

The facility should accommodate both people and wafers. The ambient air should be highly filtered; temperature and humidity should be controlled. Static charge elimination should be employed where its utility can be documented. Input chemicals must be quality checked, and point-of-use filtration is essential. The physical layout of the factory should be directed toward isolation of the wafer from particle sources in the environment. The wafer should be exposed to clean air with no more than 10 particles/cubic ft of 0.3 microns or larger. The time of exposure should be controlled.

Surrounding the wafer ambient, areas of class 100 should be designed and maintained for humans. The next larger area might be class 1000, to house that portion of the equipment which does not interact directly with the wafer. Equipment which protrudes through the clean room wall is a good choice as it occupies mimimum clean room space. Equipment layout should be modular in nature. Like equipment should be grouped together in a common area. This would allow evolutionary change to occur without disrupting the complete factory, and in conjunction with an automated material handling system, would localize the need for human movement.

Equipment must not add particles to wafers during processing. Equipment manufacturers should be prepared to demonstrate clean operation. The equipment should be capable of cassette-to-cassette operation with loading located at the SEMI-STANDARD position. Short sequences of equipment which can be linked logically, should be linked. If possible, maintenance access should be located away from the wafer input-output position on the equipment. The equipment must be reliable, i.e., it should not fail during the processing of wafers, as the loss of wafers can be very expensive. Uptime and mean-time-to-repair are less important than mean-time-to-fail. Fast fixes are a plus, but no failures are better. Maintenance schedules coordinated with mean-time-to-fail estimates can eliminate the majority of failures during processing.

Communication to the equipment to download process recipes and to report results should be available and should conform to the SECS I and II Standards. Specific messages unique to the equipment should be defined by the manufacturers and implemented in accordance with the Standards. It is understood that several Japanese equipment manufacturers are adopting the SECS Standards. They will do this well, and if domestic manufacturers do not conform quickly, they will not be in a particularly advantageous position. Process capability has been the key thrust of equipment to this point in time. That thrust must continue, but to it must be added manufacturing efficiency, and that means cleanliness, reliability and implementation of the SECS Standards.

The number of operators in the wafer fabrication facility should be minimized. They should be well trained in the activities crucial to the manufacture of products. They must apply their skills without direct interaction with the wafers. And, to minimize cost, they must apply their skills to tasks that require skill, not to tasks that can be done as well or better by automation. At least three tasks performed by operators can be done by machine activity. These are:

1. Material movement
2. Machine loading and unloading
3. Lot travel reporting.

It is estimated that one third of operator activity is involved in these tasks. They not only can be done by machine, but they can be done more accurately and reproducibly. This would improve process yield by reducing error, and productivity by applying labor only where it is required. It would

improve probe yield by reducing the interaction of people with wafers, because machines can be cleaner than people. In addition, it would improve cycle time and equipment utilization, and reduce inventory. These improvements will evolve gradually as automated material movement, automated machine loading, and physical location and tracking become part of the semiconductor factory profile. The trend toward factory automation will continue. The remaining question is, what type of material handling automation?

Material Handling Choices

Three general types of material handling automation are now being discussed. They are:

1. Single slice movement with direct process equipment input (fixed track single slice).
2. Fixed track cassette movement with fixed robotics loading cassette-to-cassette process machines (fixed track cassette).
3. Vehicle-based cassette movement with on-board robotics loading cassette-to-cassette process machines (flexible cassette).

Each of the above can provide material movement, machine loading, and material tracking. The question is, what are the relative costs and benefits of each method? The single slice method has been mentioned already as a good approach for machines which are logically linked; e.g., coat-bake-print. This method becomes expensive and output falls when too many machines are linked.

The fixed track method, with fixed robotics for machine loading, has two major drawbacks. The first is the cost of fixed robotics, considering that these pieces of equipment may be utilized less than 3% of the time. Note that a process machine which processes 50 wafers/hour requires only two cassettes per hour. If the cassette load cycle is one minute long, the fixed robotics are in use a total of 2/60 of the time (about 3.3%). However, this inefficiency is not the major drawback of fixed track material handling. That drawback is related to the continuing change discussed above.

Many different process sequences are run in a typical wafer fabrication factory. Each of these changes as process engineers modify the process to improve yield, and as equipment improves. If the process sequence changes or the equipment changes, then the fixed track must change. While this can be done, it involves structural changes in the factory. These changes involve time, expense, and additional personnel in the clean room, and as a result fixed track systems are not the optimum solution.

The approach that appears most useful and is under development is one which allows process sequence change and equipment change to occur without requiring change in the material handling system, and one which is compatible with people in the factory, i.e., flexible automation. The system utilizes automated, guided vehicles under computer control to move material as required by process flow and equipment availability. The vehicles have on-board robotics which handle cassettes, cassette boxes or "clean capsules." The "clean capsule" is designed to hold cassettes and to work with the automated equipment. It can be opened and unloaded by robotics or people, and is constructed of materials which minimize static charge and particles. A permanent bar code allows rapid identification of the boxes in intelligent bins. The robotics are capable of loading or unloading cassette-to-cassette process machines, which are linked to an equipment computer using SECS I and II.

The automated guided vehicles are battery powered and follow a taped or painted stripe on the floor. This approach overcomes the major drawbacks associated with fixed track equipment. Process sequence changes are dealt with automatically when the process flow is changed. Changes in equipment are taken care of by changing the position of the guidepath on the floor. The fact that the machine-loading robotics are on a movable platform improves their utilization. The vehicles move from process machine to process machine, which allows the on-board robot to spend more time working and less time waiting. Additionally, the system provides intelligent bins for work-in-progress storage. These bins have the capability to identify either the cassette boxes or clean capsules stored in them. This permits rapid determination of the physical location of all lots in the factory. The intelligent bins are in all vehicles and work-in-progress stations which store lots between process machines.

Material handling system products are designed to work in a clean room with people, without being dependent on them. The system is designed in a manner that allows implementation in stages, so that users can learn how to optimize its operation in their facility and minimize the emotional impact of automation. The modular nature of the equipment allows a system to be configured for most wafer fabrication areas. A major restriction is the aisle width, about two feet, within which the vehicle travels.

As already stated, yield is the issue, and the control of particles the most crucial element. Yield can be improved by reducing the overall population of the area, and, by eliminating direct human contact with cassettes of wafers, lowering wafer particulate contamination. Yield can be improved additionally by decreasing cycle time and by use of the clean capsule. The ability to move material on a need basis will improve cycle time and also will reduce inventory and its attendent cost. In addition, automated material handling will improve productivity by eliminating the need for people to do the simple tasks of material movement, machine loading, and lot tracking. All the capabilities described are possible within the framework of a system which is designed to work in the changing environment of a wafer fabrication facility.

Economic Justification of Material Handling Automation

If the implementation of an automated material handling system in a wafer fabrication factory has the potential to improve yield, productivity and cycle time, then it should be possible to estimate the magnitude of the improvement and to determine payback. The economic justification of wafer fabrication material handling automation can be based on the cost of the equipment, and the impact of the equipment

on three parameters critical to the operation of a wafer fabrication factory. These critical parameters are:

1. Labor productivity
2. Yield, both process and probe
3. Cycle time, and its impact on inventory and probe yield.

The calculation of the value of these parameters can be done in an overview fashion and compared to the equipment cost to determine payback. A wafer fabrication factory will be modeled; the required material handling automation will be added to it; and the yield, productivity and inventory gains will be calculated and compared to the cost of the automation.

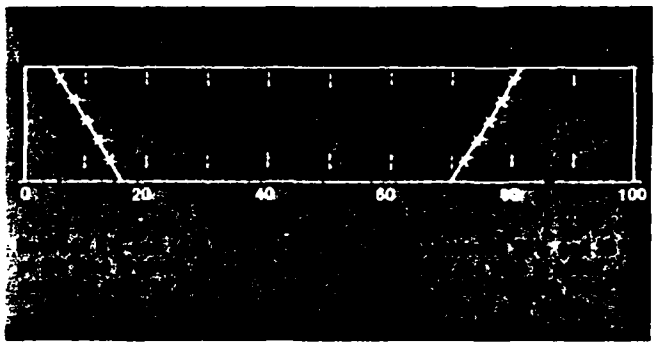


Fig. 2—Operator time distribution.

Initial Model Values

The factory model has the following characteristics:

1. 5000 wafer starts per week, 5 day operation
2. 125 mm wafers, 4.8 inch usable diameter
3. 30000 square mils per chip, 600 chips/wafer
4. 80% line yield
   2% loss due to handling breakage
   4% loss due to mistrouting
   14% loss undefined
5. 40% probe yield
6. 24 day cycle time
7. 280 direct operators
   $12 per hour: labor and benefits cost
   180 hours/month, $605K per month
8. 70 indirect operators
   $3000/month: labor and benefits cost
   $210K per month
9. $36M equipment investment
   $750K depreciation/month
10. $15M plant investment
    $125K depreciation/month
11. Finished wafer to probe cost of $200.

Table II—Cost Per Chip Breakdown

	Cost ($K/M)	(%)	Flat Cost ($)	Yielded Cost ($)
Labor, Benefits		23	37	46
Direct	605	17	27	34
Indirect	210	6	10	12
Materials	1540	44	70	88
Depreciation	875	25	40	50
Utilities & other overhead	290	8	13	16
Total	3520	100	160	200

$200 yielded wafer cost, 240 good chips per wafer, $0 83 per chip (excludes test cost).

Using the date provided by Table II as a starting point material handling automation will be applied and the impact on productivity, yield, and inventory will be calculated. The approach will assume that a constant number of good chips is produced, which will keep fixed costs constant and reduce the number of slice starts. This will allow yield improvement to generate capacity, which will not be included in the assessment of gains.

Productivity

An analysis of the work that operators do in wafer fabrication factories reveals that a significant portion of their activity is devoted to tasks that are amenable to automated material handling and lot tracking. Figure 2 shows the findings of this analysis.

The remaining time, between 74% and 56%, is spent doing operations that require skill, knowledge and decision making. The potential of an automated material handling system is that it could perform 26% to 44% of the operations now being performed by direct labor. For this analysis the midpoint of the two activities, 10% for material and 24% for load/unload will be used as the value in the cost model. Thus, 34% of labor could be replaced by a material handling system. In the model of a factory with 280 operators, a 34% reduction would involve 95 direct labor people. Of these, 28 would be accounted for by material handling and 67 by machine loading and physical lot tracking.

With the cost of labor and benefits at $12/hour, the introduction of the automated material handling system would achieve the labor savings shown in Table III. Furthermore, the system would continue to run at full capacity. An additional six indirect supervision personnel could be eliminated when 95 operators are removed, reducing cost an additional $216K.

Table III—Annual Labor Cost Reduction

Material Handling	28 Operators	$ 699K
Machine Loading	67 Operators	$ 1672K
Indirects	6 Supervisors	$ 216K
Total	103 Personnel	$ 2587K

Yield

The material handling system will improve process line yield by eliminating misrouting and a portion of the wafer breakage. Assuming that one-half of the slice breakage occurs in material handling, a 1% gain is achieved in that area. The 4% misrouting loss likewise will be eliminated, for a total gain of 5%. This raises the process line yield from 80% to 85%.

The impact on probe yield is a little more tedious to calculate. The calculation is based on the fact that chips are rejected at probe because of defects. These defects are due partially to people in the wafer environment, and the reduction of people will improve yield. The simplest relationships will be used to relate yield to people-related particulate contamination and then the impact of reducing the people-related particles will be calculated in a semi-rigorous manner.

1. $D_{Total} = D_{Particles} + D_{Other}$
   Particles are responsible for 60% of the defects;
   $D_{Particles} = 0.6D_T$
   $D_{Other} = 0.4D_T$
2. $D_{Particles} = D_{People} + D_{equip,\ fac,\ proc}$
   People generate 40% of the particles;
   $D_{people} = 0.24D_{Total}$
   $D_{equip,\ fac,\ proc} = 0.36D_{Total}$
   Thus at time 0, (baseline), the total defect density is:
3. $(D_T)_o = (D_{people})_o + 0.76\ (D_T)_o$ and, at time n
4. $(D_T)_n = (D_{people})_n + 0.76\ (D_T)_o$
   Note that D_{other} and $D_{eq,\ fac,\ pr}$ are considered constant.
   Yield has been related to defect density by several equations. The simple form below will be used.
5. $Y_o = \exp[-nA(D_T)_o]$; where
   n = number of mask levels
   A = chip size
   $(D_T)_o$ = total defect density at time 0 (baseline) per level

The same equation with subscript n applies to the case at time n, which is the time when the material handling was implemented.

The relationship Y_n/Y_o will be developed now to determine the improvement expected by the implementation.

6. $Y_n/Y_o = \exp[nA[(D_{pro})_o - (D_{pro})_n]]$
   P now is defined as the fraction of people-related defects remaining after implementation. Then:
7. $(D_{people})_n = P(D_{people})_o$
   and, since $(D_{people})_o = 0.24\ (D_{total})_o$, the following can be developed:
8. $Y_n/Y_o = \exp[nA\ (D_{pro})_o\ (1\text{-}P)] = \exp[-.24\ (1\text{-}P)\ \ln\ Y_o]$

This equation allows the calculation of yield improvement based on the value of P and the yield at time 0. Figure 3 shows the family of curves produced by this equation. Note that if $P = 1$, no yield gain can be calculated. In addition, when Y_o is low, large gains in yield can be expected by the reduction of people-generated particles.

The baseline yield for this factory is 40%; therefore the task is to estimate the factor P, which is the fraction of people-related particulate contamination remaining after implementation of the material handling system.

According to the productivity analysis, only 66% of the people remain; therefore only 66% of the particles are present. This suggests that the upper limit of P is 0.66. A more aggressive stance states that people interact with wafers only during the machine load operation. This operation has been assumed by the machine loading equipment and therefore the people interaction has been reduced to zero, i.e. $P = 0$. One reference [2] suggests that machines produce 1/10 of the particles produced by people, and that $P = 0$ would be correct.

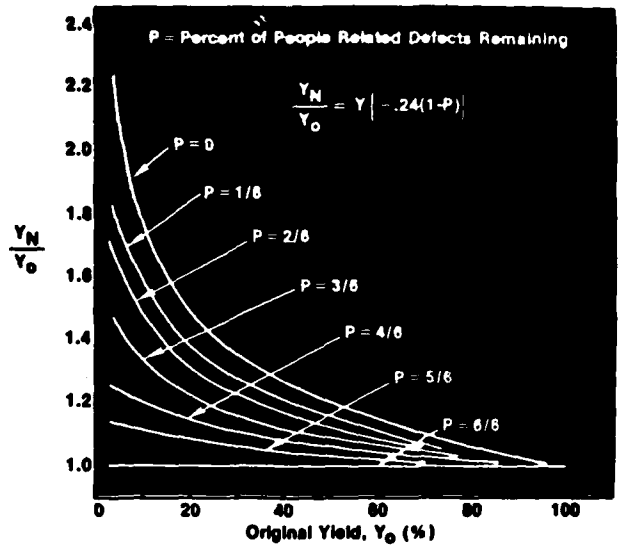


Fig. 3—Impact on yield of reducing people related defects.

Our calculation will use a conservative estimate, $P = .50$, to minimize the impact. Y_n/Y_o may then be determined and the value is found to be 1.116. This raises the probe yield from 40% to 44.6%. This benefit is attributed to the machine loading aspects of the system.

The impact of this yield gain will be calculated after a discussion of cycle time, in which some yield issues will be discussed.

Inventory Reduction

The impact of material handling automation on cycle time now will be estimated. At this point in time the estimate is based on experience with linked single slice processing equipment. Such experience indicates that a 50% reduction in cycle time is attained when linking is accomplished. This reduction now will be applied to a factory in steady state, and the resulting decrease in work-in-progress inventory will be calculated.

Many semiconductor manufacturing processes have a theoretical cycle time of four days. This implies that if 1000 wafers are started each day, (1000) × (line yield) would be output on day five. Work-in-progress would be (4000) × (line yield/day).

In actual practice, cycle times may be much longer [10], as much as 10 times longer. Thus, work-in-progress may be as large as 40,000 wafers × (line yield). This calculation will assume an actual cycle time of six times theoretical, and a reduction to three times theoretical by implementation of the material handling system.

Another important factor is at work in the reduction of cycle time. Yield increases when cycle time is reduced. Various estimates of this improvement have been up to 0.2% per day. We shall use an improvement of 0.1% per day as a conservative value. A twelve day cycle time reduction leads to a gain of 1.2% in probe yield. This raises the probe yield from 44.6 to 45.8%. The output of good chips then becomes 275/wafer, or 14.6% higher than the original value of 240/wafer.

Table IV—Original vs. Yield-Improved Factory

	Original	Improved
Starts/Week (Wafers)	5000	4107
Process Line Yield (%)	80	85
Outs/Week (Wafers)	4000	3490
Probe Yield (%)	40.0	45.8
Good Chips/Wafer (#)	240	275
Good Chips/Week (K)	960	960

Table V—Costs in Original vs. Yield-Improved Factory

	Original		Improved	
	Total ($K/M)	Yielded ($/Wafer)	Total ($K/M)	Yielded ($/Wafer)
Labor, Benefits		46		34
Direct	605	34	333	22
Indirect	1540	88	1265	81
Materials-				
Depreciation	875	50	950*	61
Utilities & other overhead	290	16	290	19
Total	3520	200	3030	195

*Depreciation has risen 75K/month due to the cost of the material handling system. This is discussed below.

Table VI—Annual and Monthly Gains for Improved Factory

Cost Gains	Annual ($K)	Monthly ($K)
Yield	3.18	265
Productivity	2.59	216
Inventory Reduction	.15	12
Total	5.92	493

If the probe yield increases from 40% to 45.8% only 3490 wafers to probe are required. The line yield increase from 80% to 85% requires that 4107 wafer starts per week be made. This allows wafer starts to decrease by 17.9% while maintaining the same output of good chips. The 17.9% decrease in wafer starts has an impact on labor and materials required, as well as on the value of the inventory.

We can see the impact of this in Fig. 3. Note that the number of input slices is now lower because of the improvement in yield. If the inventory is valued at 50% of the finished wafer cost, the original case has 21,600 wafers in process at a value of $100 per wafer (½ the finished cost of $200), or an inventory value of $2.16M.

The improved case has only 9500 wafers in process, a significant reduction. However, the value of these wafers is reduced also by material and labor cost reduction. The savings in direct labor was calculated to be $198K/month. When six supervisory personnel at $3K/month are entered into the calculations, a total cost reduction of $216K/month is obtained. The reductions in variable costs work together to lower the yielded wafer cost to $190. To this calculated cost $5 must be added to account for the increased depreciation due to the installation of the material handling system. This raises yielded slice cost to $195. The value of the inventory is then $(½) (195) (9,500) or $925K, a reduction of $1,235K. This number represents an unnecessary investment in work-in-progress. If this non-investment is valued at the interest rate of 12% then an annual savings of $148K is realized, along with the one-time savings of $1,235K.

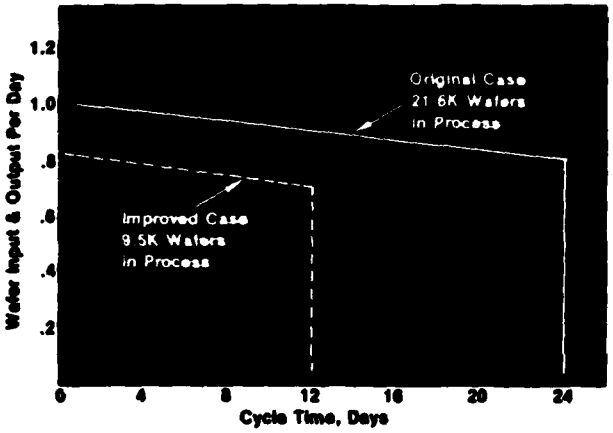


Fig. 4—Impact of material handling automation on cycle time and inventory.

Prior to the calculation of savings due to yield improvement, it should be reiterated that no credit will be given for the fact that a 5000 wafer start factory is running at only 4107 wafers/week. This represents a significant positive factor in any situation where demand for the product is growing. In fact, due to the manner in which we have made the calculation, this capacity can be exploited at the new variable cost. Thus, the potential 893 wafer starts/week would produce 877K good chips/month at a cost of $0.42/chip. This is an excellent marginal cost. In a period where demand is growing, this capacity may be essential for maintaining market share. New capacity requires up to two years to attain, and yield improvement is the quickest and least expensive method of attaining capacity.

Impact of Yield Improvement

Table IV summarizes the improved factory with these yield improvements in place, and compares it to the original.

The impact of automated material handling now can be assessed. Completed wafers to probe cost $196 and yield 278 good chips, resulting in a chip cost of $0.70. Compared to the $0.83 original cost, this $0.13 cost reduction becomes $6.5M on a volume of 50M chips/year. The new costs are compared to the original in Table V.

Summary of Gains

The factory originally was producing 50M chips per year at a cost of $42M, with an average chip cost of $0.83. The factory improved by automated material handling has the potential of saving $2,371K in direct labor, $216K in indirect cost, and $156K in inventory reduction. This is a cost reduction of $2,743K. However, the major impact derives from the improvement in yield, and consists of a savings of $4080K per year or $340K per month. However, the system

depreciation must be subtracted. When this is done the saving is reduced to $3.18M/year or $265K/month for the saving due to yield improvement. These gains are summarized in Table VI.

If the factory is orginally sized to include the gains from material handling, then the depreciation, too, can be reduced. The yield gain comes from three sources: the reduction of people-related particles, the decrease in cycle time, and the improved process line yield due to computer-controlled routing and decreased breakage.

The factory modeled above is large, but the analysis can be made for factories of differing sizes and starting conditions. The general results remain the same, but the exact values can vary considerably.

Material Handling System Cost

The cost of the material handling system which can provide the gains delineated depends on many factors. Among these are:

1. Factory capacity
2. Processes to be run
3. Number of lots in process
4. Layout of the factory
5. Time to handle material
6. Time to communicate with computer system
7. Philosophy of lot movement: fixed or discretionary
8. Amount of existing automation.

The system employed for material handling is based on automated guided vehicles with on-board robotics; work-in-progress stations with code readers for lot identification; and a computer system for control and for communication with the factory control system. The system is modular in nature and can be installed in phases. Six phases of installation will be defined. The first three phases will provide material distribution to process tunnels on a demand basis, and will include the computer control system, five distribution vehicles and 300 bins for lot storage. The second set of three phases will supply the machine-loading vehicles and 200 additional storage bins. The cost of this equipment would be priced in the range indicated in Table VII.

It should be mentioned again that this complete system cost is approximate, and also that it pertains to a large factory. It is apparent that an installation of this kind, costing $4M and having the potential to return $6M/year, is a good investment. However, some attempt should be made to match benefits to cost for each phase. The benefits of productivity can be allocated to each phase and a benefits/cost ratio can be constructed for each phase.

The productivity gain has been broken down into material handling, a Phase I, II, III gain; and machine loading, a Phase IV, V, VI activity. Cycle time improvement is the result of material handling. The impact of process line yield is 40% attributable to material handling and 60% to machine loading. Probe yeild improvement from people-related-particle reduction is associated with machine loading. From these estimates, a benefits-to-cost ratio can be constructed. It appears in Table VIII.

It should be pointed out that depreciation was arbitrarily substracted from the yield benefit gain. It could have been subtracted more uniformly, but the general result would be the same. Table VIII shows that the benefits-to-cost ratio starts above one and grows in magnitude. This implies a simple payback in one year on whatever portion of the system is implemented; an excellent situation.

Summary

The method for achieving increased integration through minimum feature size reduction is economically dependent upon attaining high yield at those feature sizes. Yield can be improved significantly by reducing the number of particles derived from people. Flexible material handling automation will provide a method for eliminating the interaction between people and wafers, and therefore will contribute toward the required yield improvement. A flexible material handling system can be implemented using programmable, computer-controlled vehicles with on-board robotics which move wafer cassettes between process machines, load process machines, and track the physical location of the wafers at all times.

The advantages of flexible material handling automation are particularly important in wafer fabrication factories

Table VII—Equipment Cost

Phase	Cost ($K)	Cumulative Cost ($K)	
I	400	400	Machine Loading Lot Tracking
II	350	750	
III	400	1150	
IV	850	2000	Material Handling Lot Tracking
V	850	2850	
VI	850	3700	

Table VIII—Benefit/Cost Ratio

Phase	Cost ($K)	Benefits From Prod.	Yield	C.T.	Total ($K)	Benefits/Cost
I	400	240	40	230	510	1.28
II	750	500	90	450	1040	1.39
III	1150	760	150	700	1610	1.40
IV	2000	1370	960	700	3030	1.52
V	2850	1990	1780	700	4470	1.57
VI	3700	2600	2600	700	6000	1.60

because of the continuing evolution of both equipment and process in those factories. Changes in process sequences or in equipment, both common occurrences in this industry, make fixed track material handling systems limited in their utility. Flexible-vechicle-based automation does not suffer when equipment or process sequences change. In addition, the use of vehicles as a base for robotics will increase the utilization of the robotics and lead toward lower cost automation systems.

Flexible material handling automation will provide the wafer fabrication factory with advantages such as increased productivity, reduced cycle time, reduced inventory, improved equipment utilization and improved manufacturing control. However, the major thrust of the system is the reduction of people-related particles in the changing environment of the wafer fabrication factory. These advantages will allow flexible material handling to become a standard technique for wafer fabrication.

References

1. J. L. Fischer, "VLSI Programation in the 1980s," Texas Instruments Engineering Journal, pp. 3-10, Summer 1980.
2. T. Makimoto, H. Nagatomi, "Automation in Semiconductor Manufacturing," pp. 11-15, IEEE 1982 IEDM Conf.
3. D. Tolliver, "Contamination Control: New Dimensions in VLSI Manufacturing," *Solid State Technology*, vol. 27, no. 3, pp. 129-137, March 1984.
4. General Subject Reference, *ibid.*
5. J. M. Duffalo, J. R. Monkowski, "Particle Contamination and Device Performance," ibid., pp. 109-114.
6. M. Parikh, Comments at March 1984 SEMI Conf. on Automation, San Jose, California.
7. M. A. Accomazzo, K. L. Rubow, B. Y. H. Liu, "Ultrahigh Efficiency Membrane Filters for Semiconductor Process Gases," pp. 141-146, *Solid State Technology*, vol. 27, no. 3, pp. 141-146, March 1984.
8. T. G. O'Neill, *Semiconductor International*, pp. 49-62, November 1980.
9. S. A. Hoenig, S. Daniel, "Improved Contamination Control in Semiconductor Manufacturing Facilities," *Solid State Technology*, vol. 27, no. 3, pp. 119-123, March 1984.
10. W. R. Bottoms, J. S. Wenstrand, "Trends in Wafer Fab, and Their Driving Ecomonic Forces," *Solid State Technology*, vol. 26, no. 8, pp. 173-180, August 1983.

James G. Harper is Vice President of Operations of Veeco Integrated Automation, Inc., where he is responsible for planning, manufacturing, and engineering. Prior to joining Veeco he was associated with Texas Instruments Incorporated for 25 years, last serving as Front End Strategy Manager. He was a Texas Instruments Ph.D. Fellow and received his degree at Stanford University in 1969.

Louis G. Bailey is Vice President of Development and Engineering of Veeco Integrated Automation, Inc., where he is responsible for the development of flexible material handling products for semiconductor wafer fabrication factories. Prior to joining Veeco he was associated with Texas Instruments Incorporated for twenty-five years. His activities included development of semiconductor materials, bubble memory materials and products, and Manager of Front End Systems Engineering. Dr. Bailey received his Ph.D. at Stanford University in 1957.

TECHNICAL PAPER

EE85-128

6.3.1

Flexible Material Handling Automation for Wafer Fabrication

abstract

Veeco Integrated Automation has developed and is manufacturing products and systems to automate material handling in wafer fabrication. The products are designed to perform material handling functions between process areas, process machine loading, and material tracking. These products improve yield by reduction of people related contamination and by reducing mishandling and misrouting. Additionally, cycle time is improved by reducing material staging time and improved productivity. Products are automated, guided vehicles with on-board robotics that accomplish material handling and machine loading. Work-in-progress stations with intelligent bins provide material at all times. The system of products is interconnected and controlled by a Local Area Network (VIAlan) utilizing 8086 microcomputers and a microVAX minicomputer. It utilizes the SEMI SECS Protocols. A description of the products and how they may be configured in a factory is included. Particular emphasis is placed on staged introduction of the product into a wafer fabrication factory.

authors

JAMES G. HARPER
Vice President/General Manager
Veeco Integrated Automation
Dallas, Texas

CARL A. FIORLETTA
Vice President/Marketing
Veeco Integrated Automation

conference

FMS For Electronics
February 25-27, 1985
Cambridge, Massachusetts

index terms

Automation
Flexible Manufacturing System
Robotics
Integrated Circuits
Systems Engineering
Semiconductors

SME

1985

Society of Manufacturing Engineers • One SME Drive • P.O. Box 930
Dearborn, Michigan 48121 • Phone (313) 271-1500

INTRODUCTION

An automated material handling system for semiconductor device manufacturing without human intervention has been developed by Veeco Integrated Automation. Additionally, portions of the system are designed to be used in conjunction with people. The system is intended to automatically move material between machines which perform the actual processing of material, to load and unload those machines, and to maintain an accurate real time log of the position of the material which is being processed. The system is particularly of value when the manufacturing process is not directly serial in nature, or contains loops of flow between various machines. It also has the capability of providing movement of material when several different material flows are being run at the same time on the processing machines.

The System has been specifically designed for the manufacture of semiconductor components, such as Integrated Circuits, but has applicability to other manufacturing processes where non-serial process flows are used. In the cases of integrated circuit manufacturing, a factory may be composed of a large number of processing machines, more than one hundred, and the material to be made into Integrated Circuits must be processed through most or all of the machines. Different Integrated Circuits will require a different sequence of processing machines and many types of Integrated Circuits will be processed in a factory at the same time. This system will provide the automatic movement of the material of all of the types of Integrated Circuits simultaneously. Other industries which could benefit from this type of system are those which manufacture assembled Printed Circuit Boards, or manufacturing processes which are potentially unsafe.

At present, the movement of material and the loading of material onto machines is accomplished in a variety of ways. Typically it is accomplished by a human. This frequently results in errors in processing sequence. In some cases conveyors or automatic carts may move material about a factory on a pre-programmed sequence, but human intervention is required for the loading and unloading of the process machines as well as the record keeping of the physical location of the material. This system can provide all of these elements of the manufacturing process automatically and it is the complete system which is the subject of this paper

SYSTEM ELEMENTS

The System is composed of three major elements; a control and communication computer network, fixed material storage stations and automated guided vehicles with or without on-board robotics. Each of these system elements provide a portion of the requirements necessary in the manufacturing process. In the simplist form the functions provided by the elements of the system are;

1. AUTOMATED GUIDED VEHICLES

This portion of the system provides the automatic movement of the material to be processed from machine to machine, and between the material storage stations. Vehicles without robotics move between these stations which have robotic elements, and can provide a method of movement between humans when the system is used in conjunction with humans. One form of the vehicle with robotics can provide material movement between the fixed stations, the fixed stations having or not having robotic elements. Another form of the vehicle with robotics can move material between the fixed stations, with or without robotics, and to the processing machines. At the processing machines it can load and unload the material to be processed.

2. FIXED MATERIAL STORAGE STATIONS

These stations provide buffers to stage the material between processing step at a location near the process machines. These stations may vary in size and capacity. Also they may have robotic elements associated with them to provide the capability to load and unload the material brought to them by the automated guided vehicles. These machine may also interact with people in situations where station loading by people is desired.

3. CONTROL AND COMMUNICATIONS COMPUTER NETWORK

This portion of the System links all of the system elements together to form the complete system. The network contains a central computer, a wired method of connecting the central computer to each of the microcomputers in the fixed material storage stations, a wired method of connecting the central computer to other computers in the factory which are not a part of this System, and a method of communicating to the microcomputers housed in the vehicles. The central computer and the microcomputers housed in the fixed stations and in the vehicles contain software which provides the control and communications cabability to the System. The software runs the mechanical actions of the System, the wired and the non-wired communications parts of the System, the overall System coordination, scheduling and material routing, record keeping and information transmittal, and the interaction with other computers not a part of this System.

A factory utilizing this system will require the communications and control computer network, and one or more of the fixed stations and mobile vehicles.

SYSTEM SUB-ELEMENTS

The entire system is composed of several sub-elements, each of these sub-elements may be used once or many times in the total system. Among those sub-elements which are frequently used are the following.

1. MICROCOMPUTER

A 8086 based microcomputer and communications board is housed in all of the fixed stations and in all of the mobile vehicles. They provide local machine control and communications. Communication to other computers which are not a part of this system may be accomplished by an ETHERNET(TM) connection, or an RS 232c or RS422 connection.

2. NON-WIRED COMMUNICATION LINK

A method of communicating to the system's mobile vehicles which does not use wire is required for the vehicles to be mobile. This could be accomplished by the use of a digital radio transmission and receiving system. However a typical semiconductor factory is electrically and magnetically "noisy" and the messages to be communicated are easily interrupted by this interference. This function of transmission and reception is preferably achieved by use of an infrared optical link. The link consists of electronic circuitry and infrared transcievers. The infrared optical transciever and circuits are mounted on all mobile vehicles to provide communication with the fixed elements of the computing system. They are also mounted on the fixed stations or on the wall or ceiling of the factory and wired into the fixed stations.

A schematic diagram of the control and communications system is shown in Figure 1.

3. TERMINAL GUIDANCE

The ability of a robotic system, whether fixed or on a mobile vehicle, to place objects on machines correctly requires that a positioning system control the movements of the robotics. In simple cases, the robot itself can provide enough control to accomplish this task. However, in cases where the distances of motion are large or in cases where the initial position of the robot can vary, such as in a robot mounted on a mobile vehicle, a method of final positioning must be employed. Vision systems can be employed for this control, but they are costly and slow. A method of final positioning of the robotics has been included in the fixed stations with robotics and in the mobile vehicles with robotics. This system utilizes the fact that the positional accuracy of the vehicle is close to correct. The system then utilizes a small electronic camera to sight a target, calculate the deviation from the correct position and provide corrective control signals to the robotics to cause it to move to the correct position. This also applies to the fixed stations with robotics where large distances are involved and accuracy is required. In all cases the concept of locating a target and adjusting the robotics to the final correct position can be achieved.

The technique employed is infrared optical. The target is a simple decal, with a printed pattern, which is placed near the correct position. A simple camera, sights the target decal, and determines the deviation from the correct position. Electronic circuits are used to apply the control signals to the robotics to achieve the final correct position. For greater accuracy, this process may be repeated.

4. INTELLIGENT BIN

In order that the material to be processed is the correct material, it is necessary that it be correctly identified. This can be accomplished by the use of an appropriate coded decal placed on the material or on the container in which the material is contained. The computer system maintains identity between the code and the material. The code can then be read to determine the exact piece or lot of material. This system utilizes a code reader in every position that material is stored, these positions are called Intelligent Bins. This allows the instant recognition of the exact position of the material when it is in an Intelligent Bin. These bins are employed in the fixed stations and in the mobile vehicles to allow complete knowledge of the positon of all material in the factory. The method of reading the code on the material is a static bar code reader, which requires no motion to achieve the reading of the code. The coded decal, placed on the container or on the material is a special bar code which allows reading in two directions, so that the direction that the material is placed into the bin is not critical.

5. ROBOTIC ARM CONTROLLERS

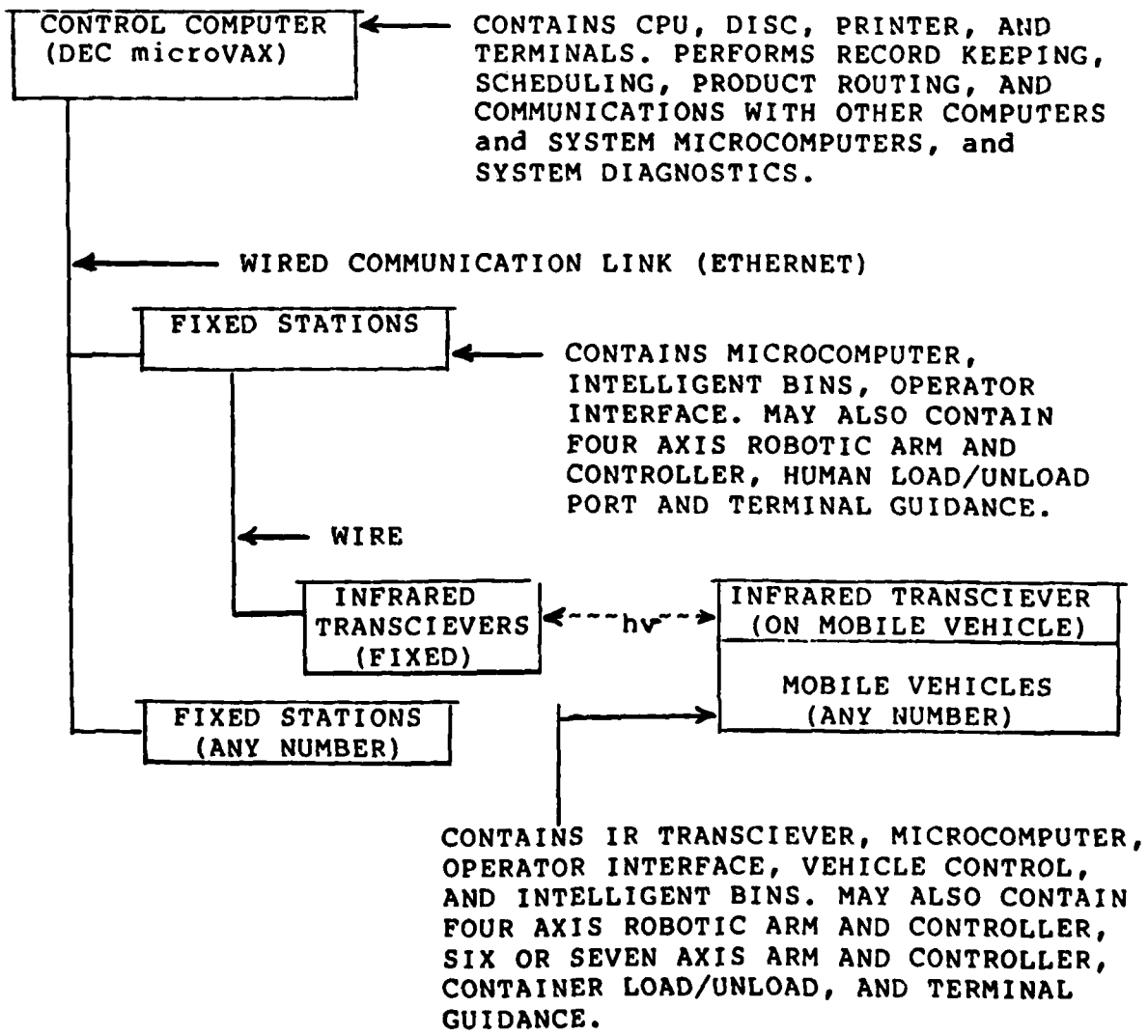
The movement and handling of the material, or containers with material in them, is achieved by the use of robotic arms. These robotic arms must be controlled to achieve the correct position to properly handle the material. Anthropomorphic arms with six or more axes of operation, such as the Unimation PUMA(TM) or the Intelledex 605 can be purchased with controllers which provide control and drive to the mechanics of the arm. Simpler arms, such as the four axis arms utilized in one version of the mobile vehicle with robotics and in the fixed station with robotics, require methods of driving and controlling the mechanics. This System utilizes standard motors, power supplies, motor drivers, and motor controllers to drive the simple arms and the standard control systems purchased with anthropamorphic arms. Control of the simple arms is achieved by the use of position sensors, motor stepping feedback, and the microcomputer to determine exact position. Using these methods a robotic arm can perform all of the functions of handling and placing material.

6. OPERATOR INTERFACE

In order that a person can transmit information to the system or can interrogate the system for information, a method of interaction between humans and the intire system must exist. Terminals linked to the system are one method this can be accomplished. This does not allow information directly at the equipment. An operator interface on each piece of equipment is part of this System. It consists of a display, a keyboard and control circuits which allow the communication with the microcomputers in all parts of the System.

Using these System sub-elements, and others to be described below, the physical elements of the System can be described.

SYSTEM OVERVIEW



Material is moved by the mobile vehicles from fixed stations to other fixed stations or to process machines, where it is loaded on the machines for processing.

Material is stored in the fixed stations between operations on the process machines if the next required process machine is busy or out of order.

The distributed computer system, both the Central Computer and the microcomputers in the vehicles and the fixed stations, keep track of all material, the position of the material, and its next required action.

SPECIFIC PRODUCTS

1. AUTOMATED GUIDED VEHICLES

Three types of automated guided vehicles are available. The vehicle functions of drive, steering, path finding, and safety are provided by the base chassis of a automated guided vehicle. This vehicle, purchased from Litton, follows a painted or taped guidepath on the floor using optical guidance. It also reads codes on the floor for positioning. All vehicles require the Infrared Communications Link. Additionally, all vehicles require the Microcomputer, the Operator Interface and the Intelligent Bin. Implementations of the automated guided vehicle portions of this total system are discussed below. In each case the new sub-system elements are noted. The implementations are:

1. Vehicle 1

This vehicle contains only the base sub-system elements noted above; the Infrared Link, the Microcomputer, the Operator Interface and the Intelligent Bins. Material must be placed onto and removed from this vehicle by other machines or by people.

2. Vehicle 2

This vehicle has all the sub-elements listed above in Vehicle 1. Additionally it has a four axis robotic arm and arm controller. The function of this arm is to transfer material or boxes containing material between the V2 and storage stations which do not have on-board automation.

A picture of the Vehicle 2 is shown in Figure 2.

3. Vehicle 3

This vehicle has all the sub-elements listed above in Vehicle 1. Additionally it has a six or seven axis robotic arm and arm controller and may have a box loader/unloader. The function of the arm is to place material onto process machines for processing operations. The Box loader/unloader performs the function of presenting the material to the robotic arm if it is housed in a box. The robotic arm is commercially available, such as the Unimation PUMA(TM) or the Intelledex 605.

2. FIXED STATIONS

There are two implementations of the fixed stations. Common elements present in each the Microcomputer, the Operator Interface, and the Intelligent Bins. The number of Intelligent Bins can vary from one to more than one hundred in each implementation as required by the application. The two major types are:

1. Local Stations

These contain the sub-elements listed above; the Microcomputer, the Operator Interface, and the Intelligent Bins. The function of this implementation is to provide storage near the process machines. The Local Stations have a small number of Intelligent Bins, from one to thirty, and no automation associated with them. They can be loaded and unloaded by the Vehicle 2 and the Vehicle 3.

A picture of the Local Station is shown in Figure 3.

2. Zone Stations

These contain the sub-elements listed above; the Microcomputer, the Operator Interface and the Intelligent Bins. Additionally, they contain a Four Axis Robotic Arm and Controller, Terminal Guidance, and an Operator Load/Unload Station. The function of the Zone Station is to provide storage for a large area of the factory. The Zone Stations have a large number of Intelligent Bins, from twenty to more than one hundred. They can be loaded and unloaded by the on-board Robotic Arm from all of the vehicles, the Vehicle 1, the Vehicle 2, and the Vehicle 3.

A picture of the Zone Station is shown in Figure 4.

The fixed stations provide the connection of the Infrared Communications Link to the system. The links are wired into the Microcomputer in the fixed stations and up to ten links may be wired into the same fixed station.

3. CONTROL AND COMMUNICATIONS COMPUTER NETWORK

The computer control and communications network is composed of a minicomputer (DIGITAL EQUIPMENT CORP. microVAX(TM)), an INTERLAN ETHERNET(TM) communications board, and an ETHERNET(TM) communications cable which attaches to the microcomputers in the fixed stations. As stated, communications to mobile vehicles is accomplished by the Infrared Link, which transmits information between the fixed stations microcomputers and the mobile vehicle microcomputers.

Links from the factory MIS to the computer system can be accomplished in at least three ways. They are;

1. RS-232c link from minicomputer to the MIS.

2. Ethernet link to the MIS

3. Fixed station Microcomputer RS-232c or 422 link to the MIS.

The preferred implementation is the ETHERNET link. Additionally, it should be noted that process machines may be connected to the fixed station microcomputers and that information may be transmitted to the factory MIS through the ETHERNET.

SUMMARY

The automated material handling components described above provide advantages to the user. Productivity is improved as is estimated that up to one-third of the personnel are involved in material distribution, machine loading, and logging operations. Inventory reductions are achieved by the reduction of material staging. This improves cycle time through manufacturing and improves product yield. Product yield is also improved by the use of automation to handle the semiconductor material. The robotic approach to material handling reduces the number of operators in the factory. Operators are a source of particulate contamination which causes yield loss. This equipment is far cleaner than people and therefore yield loss from people contamination is reduced.

The economic impact of automated material handling in semiconductor device manufacturing has been discussed in detail in the July, 1984, Solid State Technology and will not be repeated here. Future efforts will be directed toward defining the degree of cleanliness achieved by use of automated material handling.

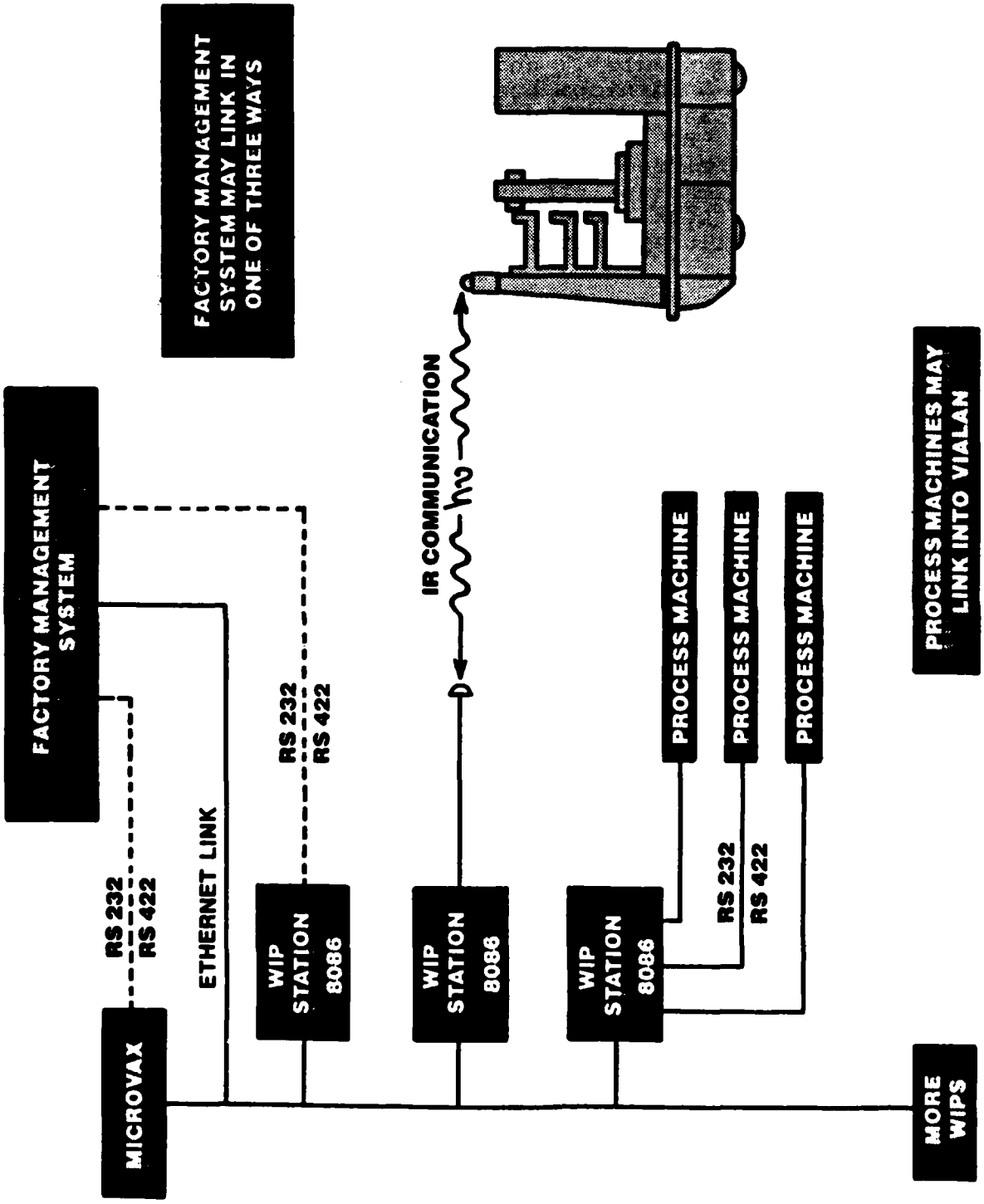


Figure 1. Schematic of Communications and Control System.

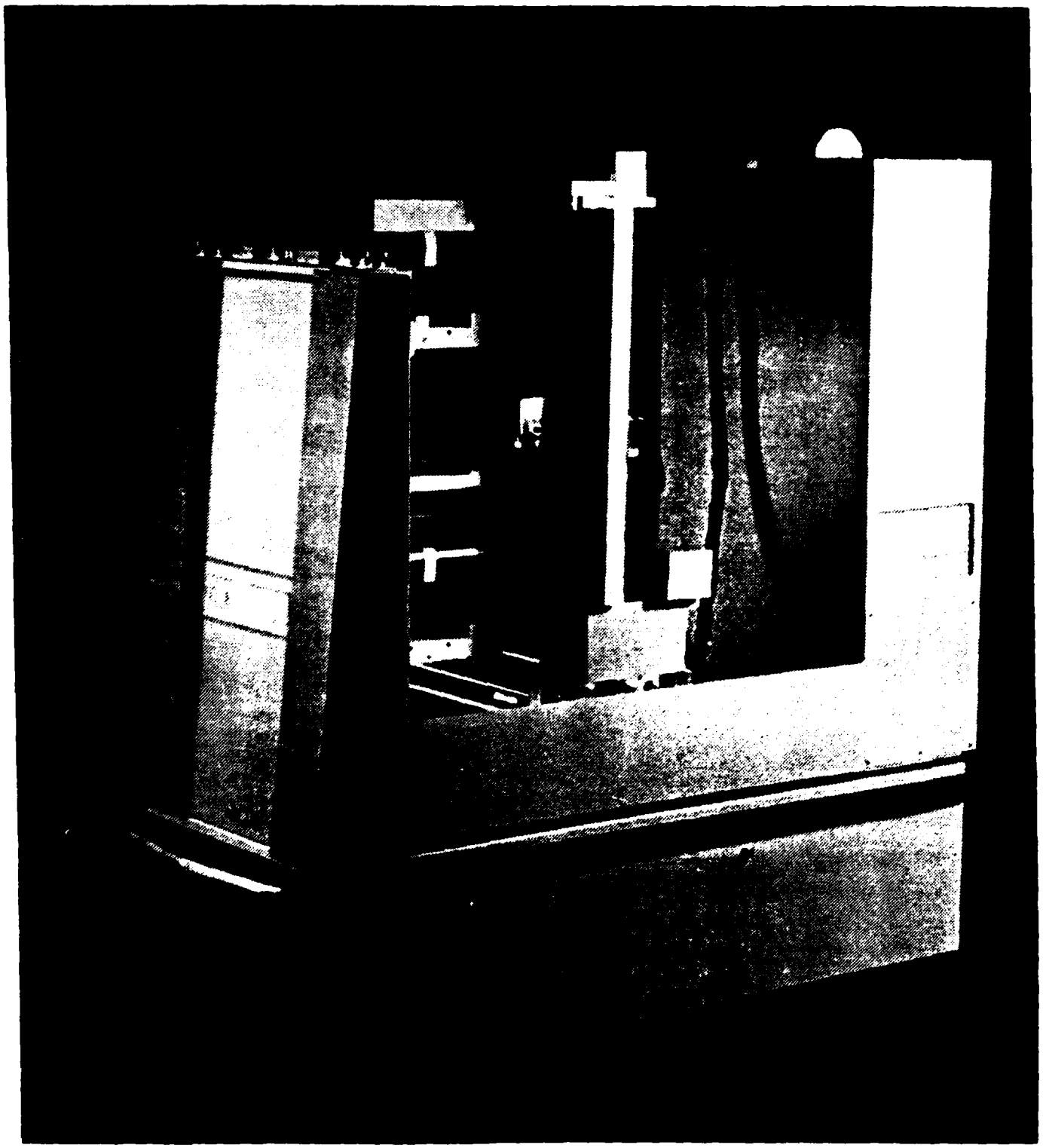


Figure 2. Vehicle 2 with Four Axis Robotic Arm.

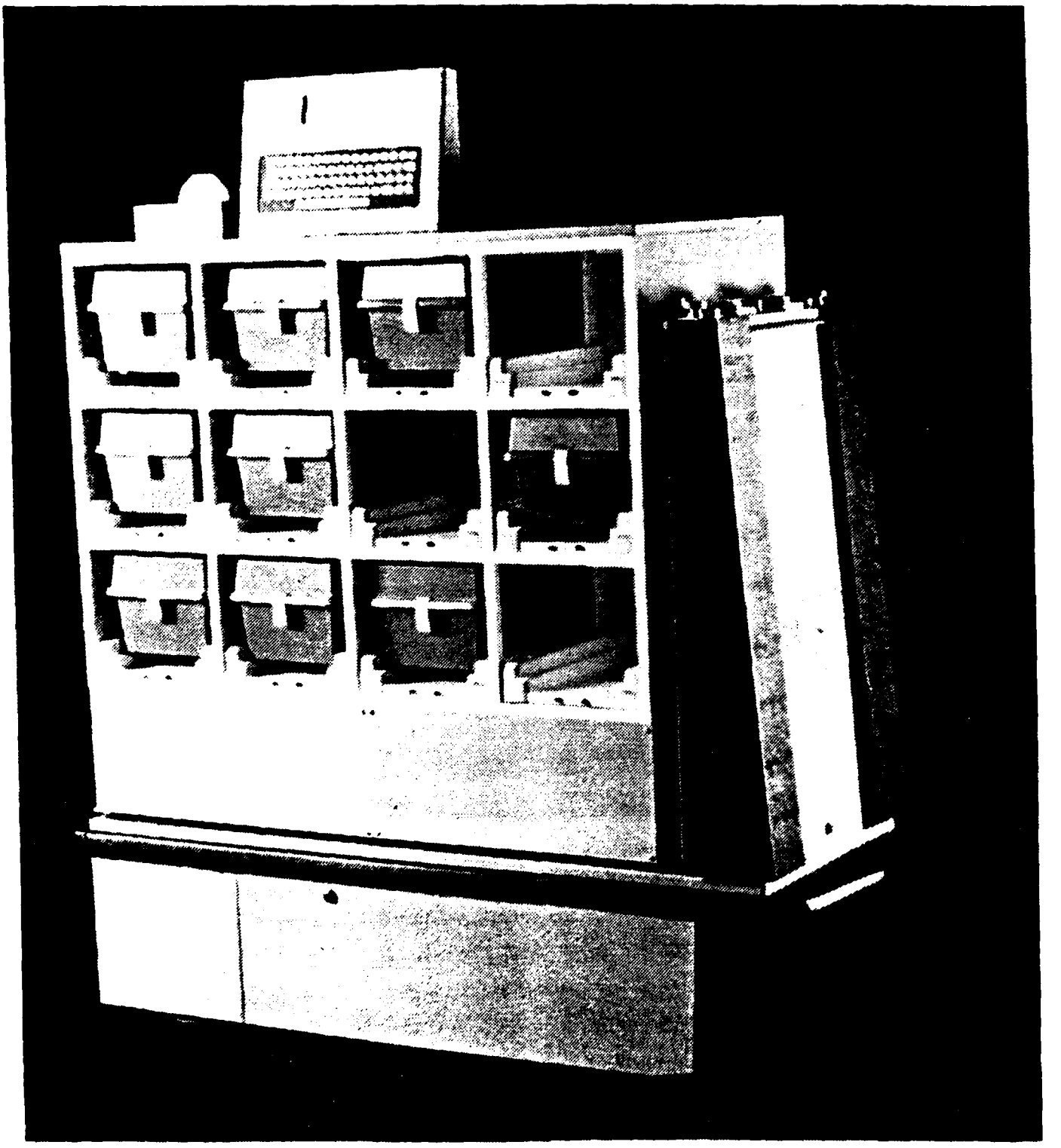


Figure 3. Vehicle 1 with Local Work-In-Process Station.

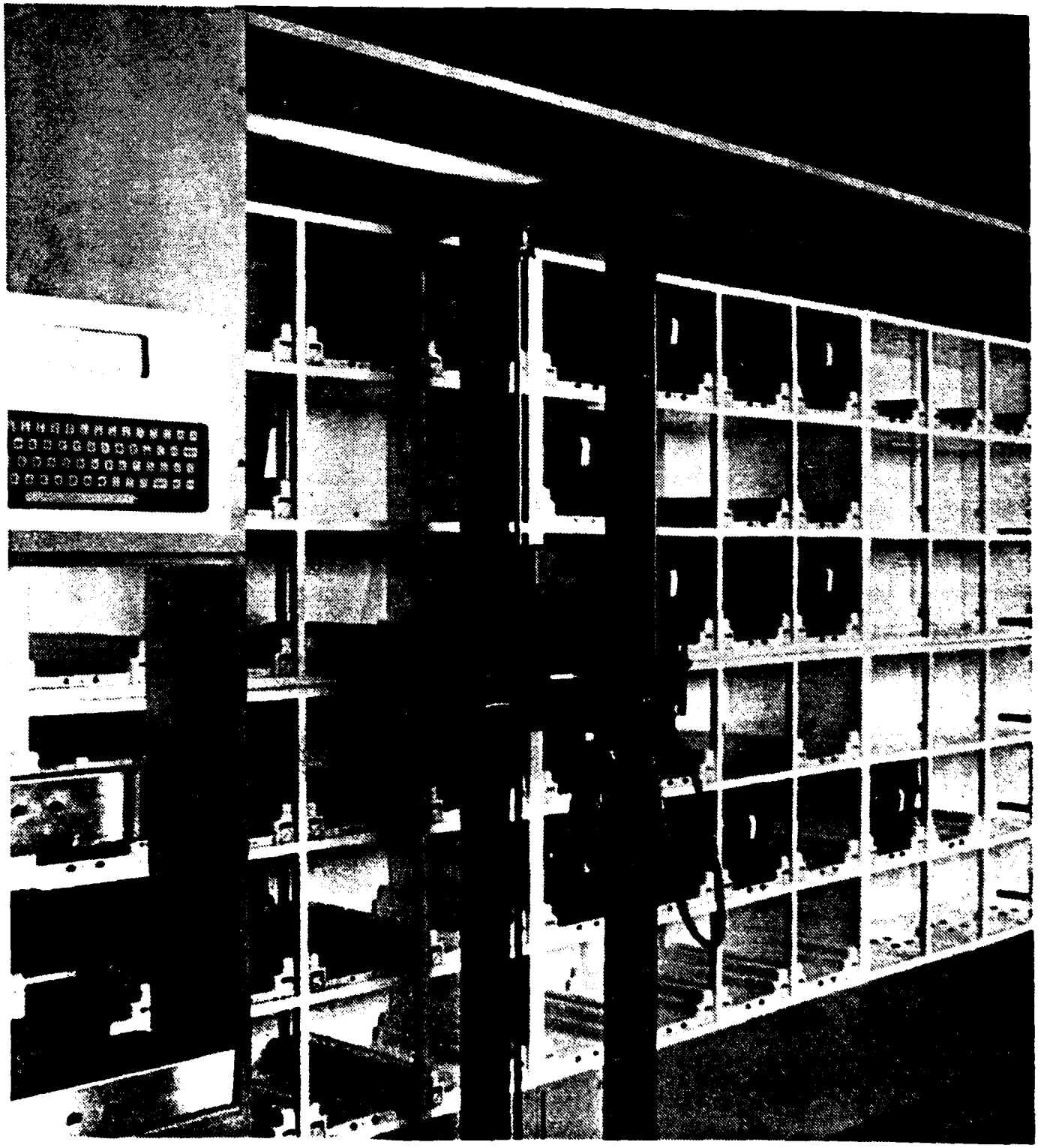
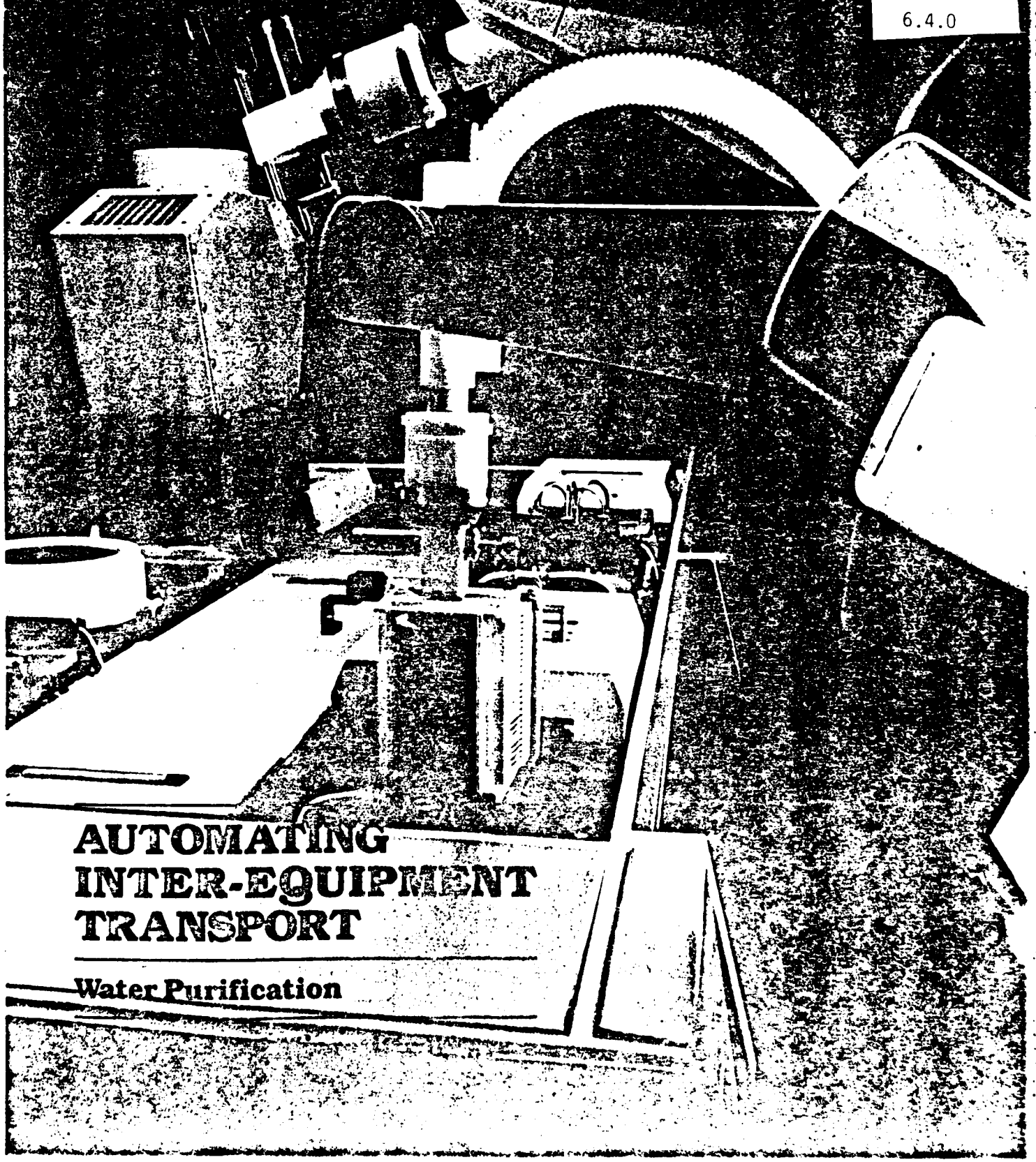


Figure 4. Zone WIP Station with Four Axis Robotic Arm.

6.3

SEMICONDUCTOR INTERNATIONAL

PROCESSING, ASSEMBLY & TESTING • MARCH 1985 • A CAHNERS PUBLICATION



AUTOMATING INTER-EQUIPMENT TRANSPORT

Water Purification

Reprinted from Semiconductor International, March 1985

Automating Inter-Equipment Transport

Cassette-to-cassette automation is currently available on most front-end equipment, yet wafer transport between equipment is only beginning to be developed.

Peter H. Singer, Associate Editor

Sitting in your control room high above the clean room, you look out over a vast expanse of white cabinets which you know to contain sophisticated wafer processing equipment. A number of vehicles, looking remarkably like electrically propelled mail-carts with robotic arms perched on top, whiz down the aisles transferring cassettes of wafers from one piece of equipment to another. Suddenly, a red light flashes on your console; a stuck valve causes an etcher to go down. No problem — within seconds you have programmed the "robots-on-wheels" to bypass the malfunctioning etcher and have rerouted the cassettes to one that can handle the extra load, and downloaded the proper process recipes to the new system. Sound far in the future? Not really — fully automated facilities such as this are technically feasible and will probably be in operation within the next two to three years.

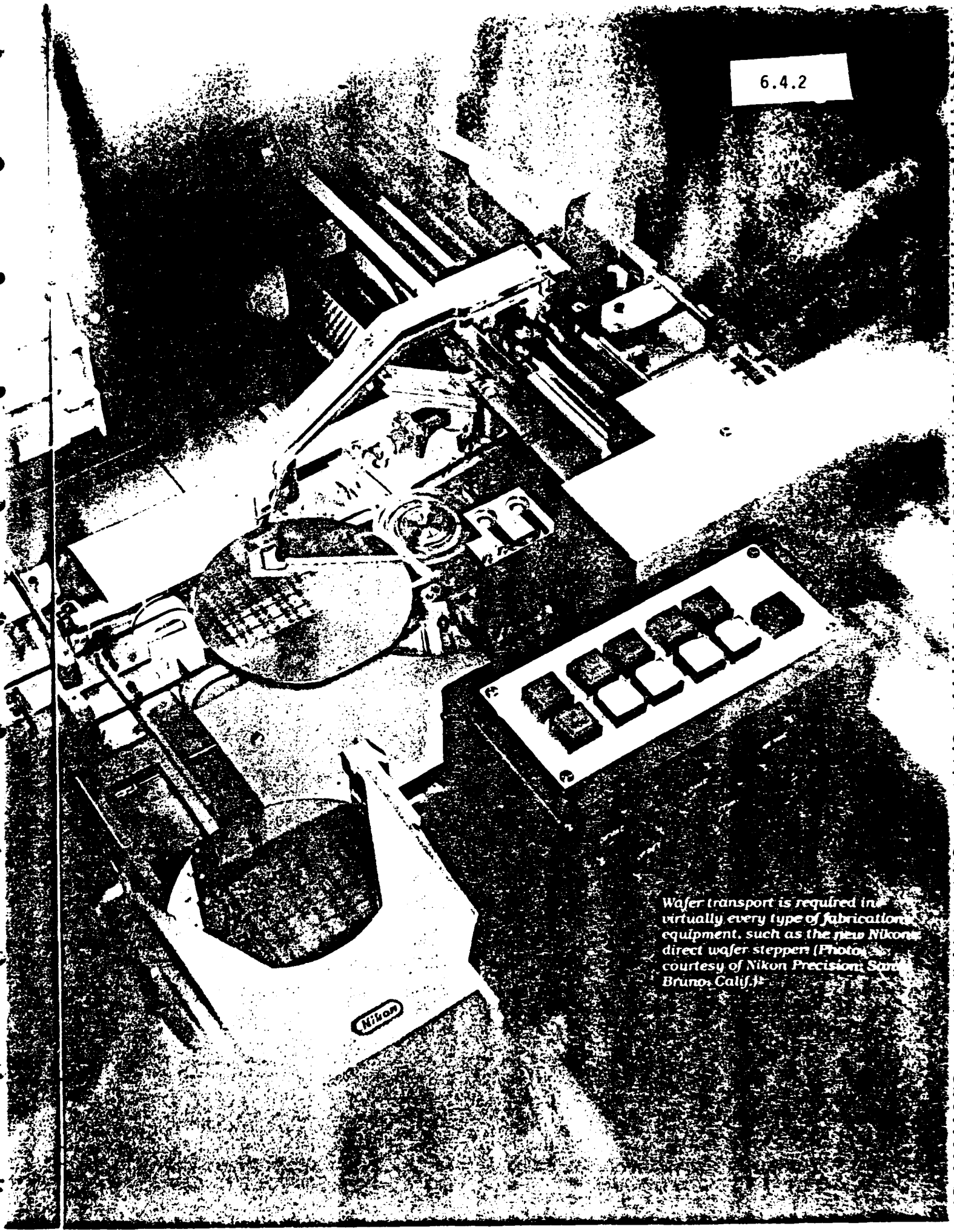
The benefits of automation are well known to most people working in the semiconductor industry, those being reduced contamination and increased productivity. The reduction in contamination is a direct result of reducing the number of people required in the clean room. Increased productivity results from better monitoring of production through the use of host computers networked to the processing equipment.

Most semiconductor manufacturers have not yet felt it to be cost-effective to make any drastic changes in terms of automating their facilities. Instead, they are gradually incorporating automated operation into existing facilities while designing it into new facilities. They have also demanded (and gotten) more automated features on new processing equipment.

Two separate issues must be addressed when attempting to implement automation; they are process control and material handling. Process control is a complex task, requiring computer control over all processing steps which, in turn, requires inter-equipment communication standards (i.e. SECS II developed by SEMI), the capability to monitor/control the necessary process parameters in each piece of equipment, and a host computer. Progress in this area has advanced rapidly and a number of companies have developed commercial computer-aided manufacturing (CAM) systems and most equipment vendors offer SECS II capability and a fair degree of the necessary process control.[1]

This article will discuss the second aspect of automation, material handling. Compared to other industries, material handling in the semiconductor industry is relatively simple since there is only one material that needs to be moved automatically — the silicon wafer. Wafers must be moved through the different fabrication processes, which may employ high temperatures and corrosive chemicals. In addition, some processes can utilize batch processing and others must use single wafer processing, requiring different handling techniques.

A variety of mechanisms have been developed by equipment manufacturers to transport wafers within such equipment, including polypropylene o-ring belts, air tracks, walking beams and stainless steel belts. In most cases, transport systems are designed and



Wafer transport is required in virtually every type of fabrication equipment, such as the new Nikon direct wafer stepper (Photo courtesy of Nikon Precision, San Bruno, Calif.)

built by the original equipment manufacturer (OEM), resulting in a fairly individualized approach for each application. Several features are required of any wafer transport system, regardless of its application. Such systems should be accurate, reliable, adaptable for different requirements, upgradable, serviceable and cost-effective. In addition, special capabilities may be required, such as programmability, wafer sensing, buffers, reject switches, a travel distance of 1 to 2 m and the ability to move around corners. Of course, the single most important feature of all is that of cleanliness, since wafer transport mechanisms are often required to operate in a Class 10 or better environment.

Polypropylene o-ring belts are probably the most commonly used method of intra-equipment transport. In this type of system, the wafer rides on "rubber bands" which are driven by a stepper motor under microprocessor control. Airbearing transport, which is licensed by GCA Corp., Bedford, Mass., utilizes a cushion of air to float and move the wafers. Flat metal plates with many small holes make up the airbearing track. Air blown through the holes makes the wafers float. If the holes are set at an angle, the air propels the wafer along the track from one module to the next. Pressurized clean dry air or nitrogen is used to float the wafers.

For harsher environments, where corrosive chemicals or high temperatures are used, it has been necessary to develop alternative transport technologies. A walking beam technique, for example, has been developed by Brooks Associates, Inc., North Billerica, Mass., called the Orbitrac. This mechanism is especially designed to minimize contamination, since no sliding or rubbing surfaces are involved. The system operates by having two parallel beams, one within the other, which move in an elliptical orbit, 180° out of phase with one another. In this manner, the wafer is "walked" along, first moved by one beam and then the next. As an alternative, many systems utilize stainless steel belts which operate on a principal similar to the o-ring belt method, but can withstand the harsh environment.

Of course, many equipment vendors have developed more unique approaches to wafer transport within their systems. In the diffusion area, for example, it has been necessary to develop a method of transferring wafers from standard cassettes to quartz boats. This is usually done by "pick-and-place" units which operate similar to a vacuum wand, contacting the wafers on the backside and individually transferring them to the boat. Units have also been developed which will transfer the whole cassette of wafers at one time.

Robotics

Pick-and-place units should not be confused with more advanced robotic systems which have a much wider range of motion and flexibility. Robotic manufacturers have recently introduced units that are compatible with clean room environments.[2]

There is only one material that needs to be moved automatically — the silicon wafer.

Robots are especially useful in batch processing systems, such as sputtering systems, where wafers are loaded onto a planar holder from a cassette. Robots are also useful for automatically moving cassettes of wafers from equipment onto a cart or onto an automated inter-equipment transport system (Fig. 1).

Inter-equipment transport

There are basically three approaches to transporting wafers from one piece of equipment to another. First, the two pieces of equipment can be directly linked together, both mechanically and electronically. Secondly, the cassettes can be transported over a fixed modular track or thirdly, the cassettes can be transported by a guided "robot-on-wheels" type of vehicle.

The direct link approach is well suited for some small groups of processing steps, such as scrub, coat, bake and align. In fact a number of photoresist processing systems, commonly called "track" systems, are currently available that combine modular processing functions within one cabinet. More recently, track system manufacturers have designed direct interfaces to lithography equipment. Another area that could logically be direct linked is between an ion implanter, a stripper and a vertical furnace since the amount of time required by each step is roughly equivalent. Randy Karl, program manager for General Signal's Automation Group, believes these areas may be directly linked in the near future, due to larger wafer size. "As we get into larger wafers, you're going to see more of a wafer to wafer process than a cassette to cassette process. The track systems are really going to come into play as we get into big wafers," said Karl.

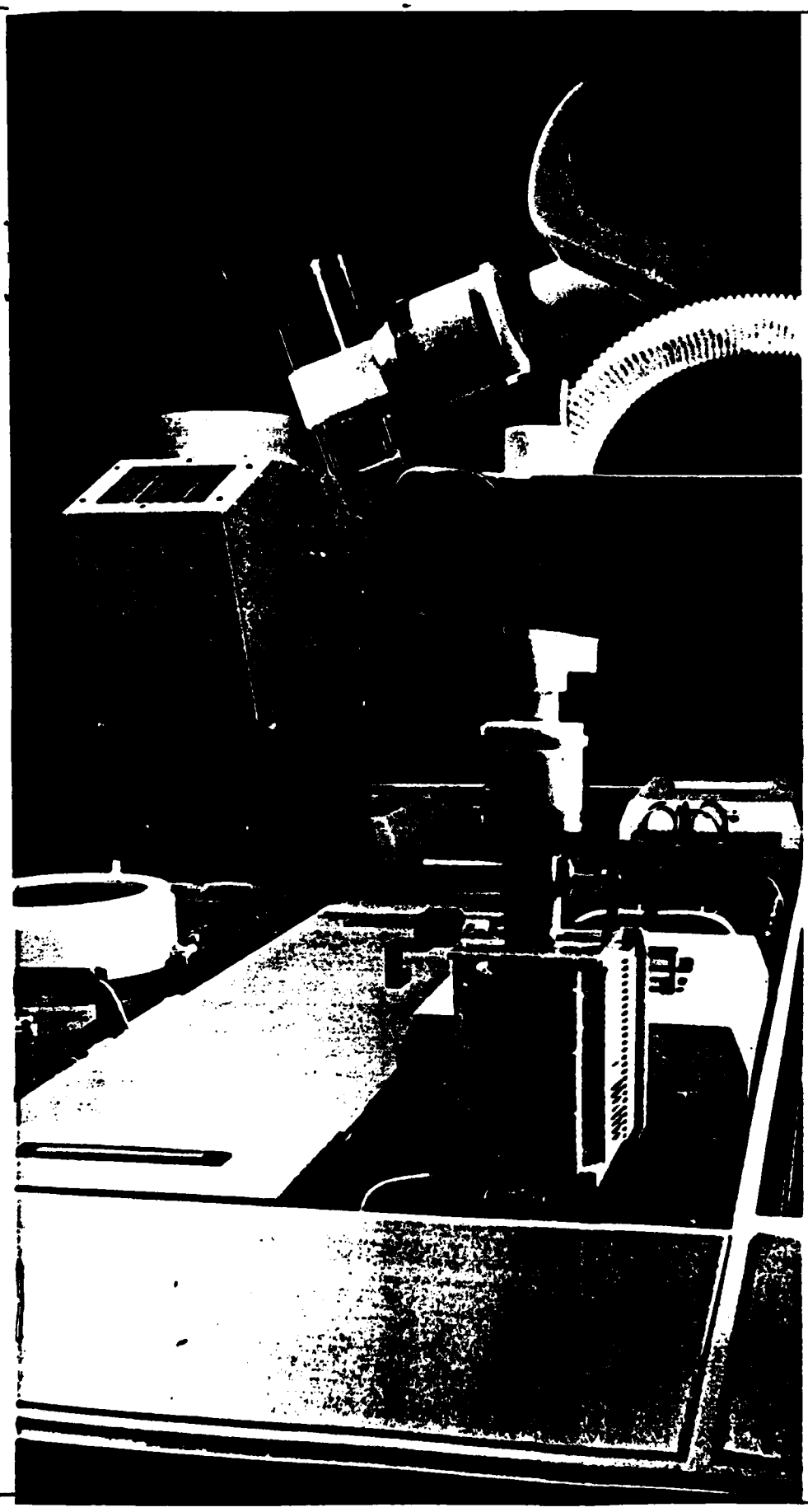
It is doubtful, however, that the direct link approach could be successfully applied to an entire fabrication line. The major problem with the direct link approach is that if one piece of equipment goes down, the entire line shuts down. There are also problems in trying to get equipment vendors to cooperate with one another in developing a mutually compatible interface. According to Jim Harper, vice president and general manager of Veeco Integrated Automation, Dallas, Tex., "If you try to extend the direct link approach to too many pieces of equipment, the downtime begins to really work against you." A

The second and third approaches to inter-equipment transport are both based on a cassette transport technique, since most equipment is already automated on a cassette-to-cassette basis. Presently, cassettes are manually carried between equipment greatly increasing the chance of contamination. The two approaches to automated inter-equipment transport are fixed track or a more flexible "robot-on-wheels" approach.

Fixed track

The only fixed track system that is currently being marketed has been developed by Nacom Industries, Inc., Tustin, Calif. Nacom offers the Namtrak wafer transport system where wafers are moved in an isolated, clean environment through sealed modular sections. By attaching straight sections and turntables together, a bi-directional or continuous loop system can be installed in a facility at work station level or elevated above doorway height. System complexity varies from simple relay logic controlled "call/send," single railcar units to fully automated systems with robotic auto-load/unload stations and multiple railcars.

According to Nacom, a Namtrak system is currently in use at AT&T's



1. Robotic arms are especially well suited for loading and unloading cassettes of wafers to and from processing equipment. Robotics are even more cost-effective when mounted on an automatically guided vehicle as shown. (Photo courtesy of Veeco Instruments, Plainview, N.Y.)

Kansas City facility and several more are on order for their Allentown plant. (Varian has developed a similar system, called the Autotrack, but has decided not to actively market it.)

One major disadvantage of the fixed track approach is the problem of loading and unloading of cassettes to and from the track system. This can be easily accomplished with a robotic arm fixed at each gateway, yet robotic arms are costly and it is usually not cost-effective to use them in this manner. Nacom, however, is currently developing an inexpensive, simple robotic arm dedicated to this purpose.

"We are developing a system that is at work level and is from process machine to process machine. Our robotic arms are very simple and very economical to manufacture," said Jim Bushong, Nacom's design engineer. The system can also be used as a clean storage cabinet, since each station elevator is capable of holding 12 cassettes. "Because it's a modular system, it is relatively easy to change a piece of equipment. It may require reprogramming, but it's usually a matter of adding modules or lengthening or shortening track sections," Bushong added. He also said the major benefit of a fixed track approach is its lower cost when compared to more expensive flexible transport systems.

Flexible transport

An alternative approach to inter-equipment automation has been developed by Flexible Manufacturing Systems (FMS), Los Gatos, Calif., scheduled to be unveiled at the Semicon West show in May of 1985. The FMS system consists of four major components: a mobile transport unit (MTU), a docking module, a central control computer and an intelligent work-in-progress (WIP) station (Fig. 2). The MTU is a vehicle that transports wafer cassettes or boxes from one work station to another by a robotic arm. The docking module is mounted on the production tools or WIP

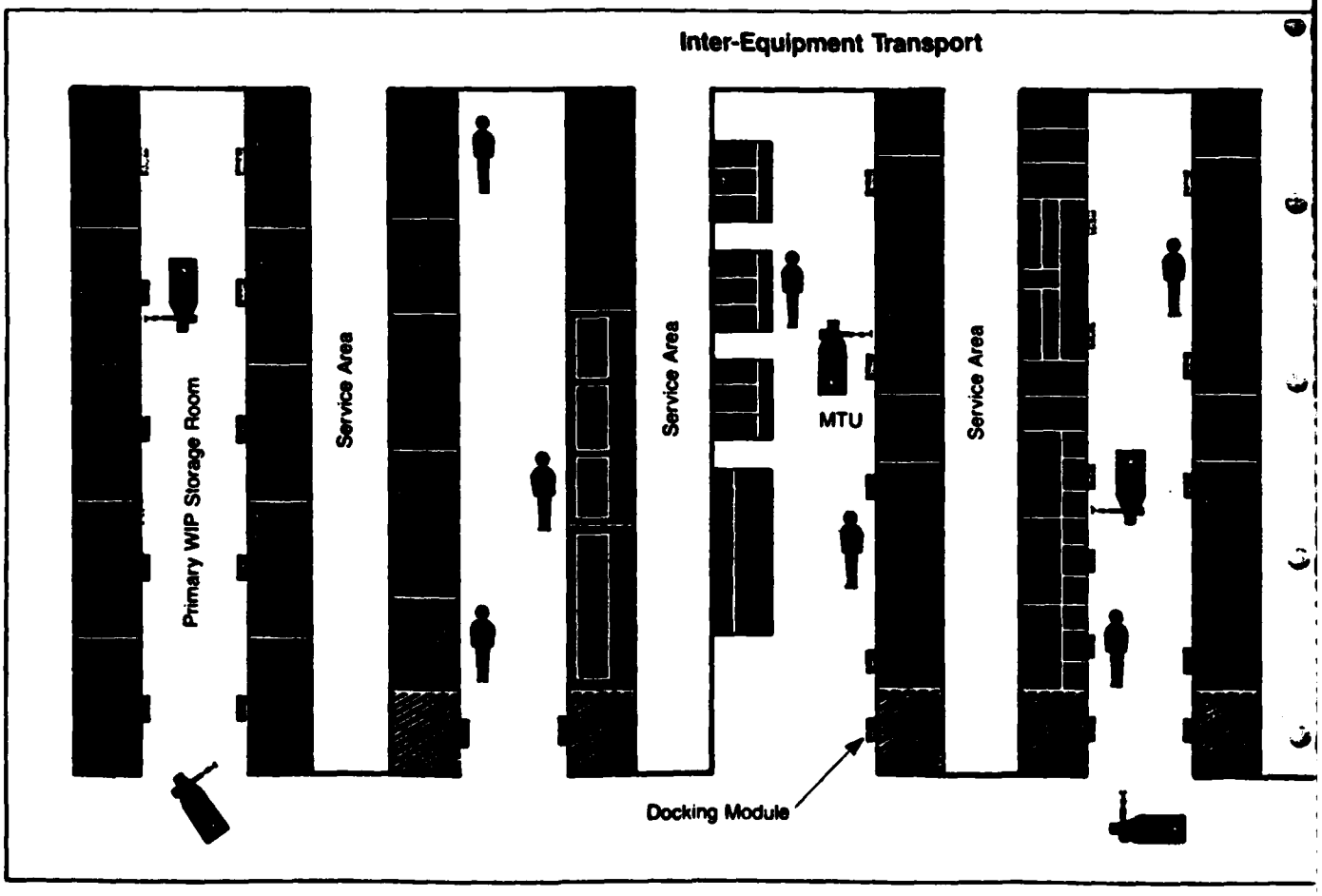
stations to provide a docking point for the MTU. The central control computer manages all MTU guidance and location monitoring functions as well as local interfacing between the MTU and production tools being serviced. The WIP station provides local storage and inventory control for 16 wafer cassettes or cassette boxes.

Veeco Instruments Inc., Integrated Automation Div., Dallas, Tex., have pioneered work in the flexible transport area, and currently offer an advanced automated wafer transport system. The Veeco system consists of automated guided vehicles with robotic capabilities, a distributed processing communications and control network, a software system which controls system components and flexible guidepaths for the automated vehicles to follow. "We feel that the typical semiconductor manufacturer will find it far easier to approach automation using a flexible approach than fixed track, particularly in an environment where he frequently changes his process and certainly would resent having to rip-up portions of the fab area," says Ed Braun, executive vice-president of Veeco. Three different types of robots are offered, the most complex of which is the Veebot III. This vehicle may receive lots from either a piece of processing equipment or a work-in-progress (WIP) station and utilizes a seven-axis robotic arm from either Intelledex or Unimation. The WIP station (Fig. 3) provides storage locations for 11 lots of either 1 or 2 cassettes per lot. The Veebot II has on-board robotics capable of loading and unloading WIP stations and the Veebot I distributes lots between WIP stations although it has no robotic capabilities. While in transfer, the cassettes are stored in standard Fluoroware boxes that have a bar code attached to the bottom of the box. "Each time the cassette is placed or removed from an intelligent bin, the bar code is read. The central computer identifies what that lot is and what its next processing step should be which also permits you to not only have on demand materials distribution, but you can have real time inventory management," says Braun.

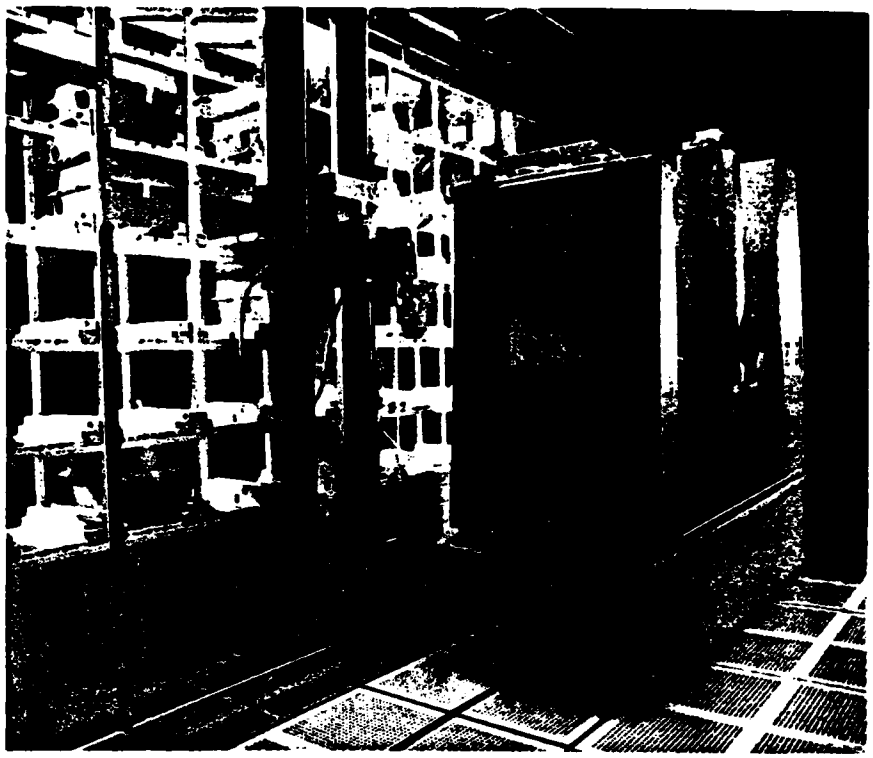
SMIF

The major benefits of automating inter-equipment transport are a reduction in contamination through reduced operator handling, an increase in productivity and better inventory control. Recently, Hewlett-Packard, Palo Alto, Calif., developed a technology which also is designed to reduce contamination and can be used as an interim step between manual and automated cassette transport. Mihir Parikh, formerly depart-

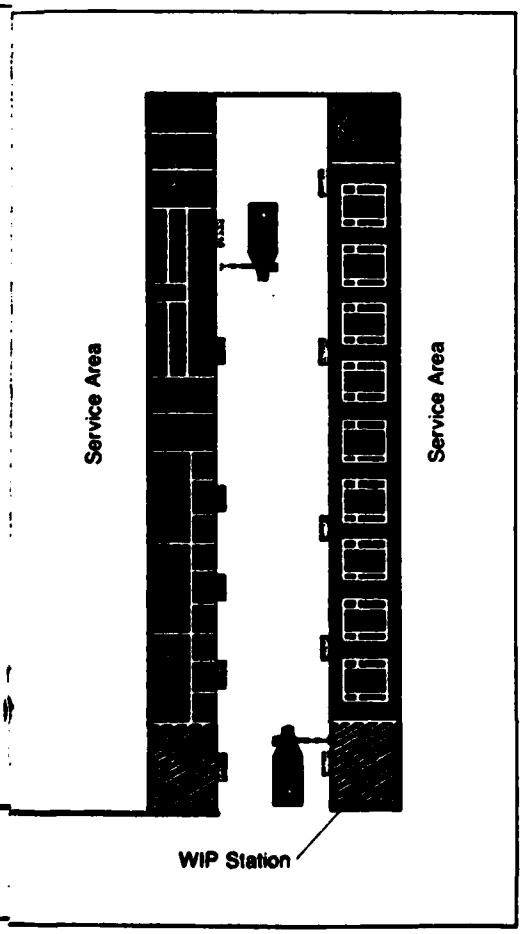
2. A robot-on-wheels approach is especially suited to the tunnel clean room concept. A docking unit on each piece of equipment allows the robot to accurately transfer cassettes. (Illustration courtesy of FMS.)



ment manager of the Process Automation Dept. at Hewlett-Packard Labs, Palo Alto, Calif., has since started a company called ASYST Technologies, Fremont, Calif., to commercially develop the technology. According to Parikh, "Human operators and technicians will be around in the fab for at least the next six years, certainly until the end of the decade. It is very trivial for them to provide the physical movement of the wafers. The SMIF philosophy is that you might as well let people transport the wafers, but bring them in a sealed container to keep them clean and then evolve into automated materials movement." Parikh added that the SMIF concept could also significantly reduce the need for clean room space. Called the Standard Mechanical Interface (SMIF) concept, the idea is that the best environment for wafers is a small volume of still, particle free air, with no internal source of particles. Ulrich Kaempf, project manager in the Process Automation Dept. at HP, describes how it should be implemented. "At one place in the fab, the wafers have to enter what we call a SMIF environment. Then, once in that environment, they would remain there as they move from equipment to equipment, either inside the SMIF box during storage, cuing, transport or underneath the SMIF canopy."



3. Work-in-progress stations are required to store wafers in between process steps, since some processes take longer than others.



One criticism of the SMIF approach has been that the boxes could collect contaminants and then deposit them on the wafers. Kaempf reports, however, that HP has not experienced any problems with this. "Particles that enter or are still in the box will rapidly fall out either to the bottom of the box or to the sidewall and cling due to electrostatic charge. We have had a SMIF box in operation for several months now and have had no need to clean out the inside," he said.

HP is currently licensing the SMIF technology for a nominal charge to any interested party. Further information can be obtained by contacting Ed Wong, Hewlett-Packard, Palo Alto, Calif. In addition, SEMI is currently evaluating the SMIF concept in developing a standard for "Protected Methods of Inter-equipment Wafer Transport." For additional information on SEMI's activities, contact Maynard Coulton, chairman of the Equipment Automation Subcommittee at Applied Materials, Santa Clara, Calif. (Further information on the SMIF concept will be presented in an upcoming issue of *Semiconductor International.*) □

References:

1 P. Singer, "Computerizing the Process Line: A Must for Automation," *Semiconductor International* (April, 1984) pp. 68-73.
2 R. Iscoff, "Robots in the Clean Room," *Semiconductor International* (November, 1984) pp. 50-57.

Reprinted from **SEMICONDUCTOR INTERNATIONAL** March 1985

SPECIAL REPORT

IC PRODUCTION LINES MOVE CLOSER TO FULL AUTOMATION

by Jerry Lyman

ROBOTS, GUIDED VEHICLES, STANDARD INTERFACES, AND COMPUTER NETWORKS POINT TO HIGHER CHIP YIELDS

Despite an ongoing recession, the integrated-circuit industry is in the midst of serious changes driven by the need to cut costs and raise yields of current and future very large-scale ICs. IC production is moving from its current phase of semiautomation toward almost full automation of the entire manufacturing process, including assembly and testing. The goal is to remove people—the major source of circuit-killing particulate matter—entirely from the process.

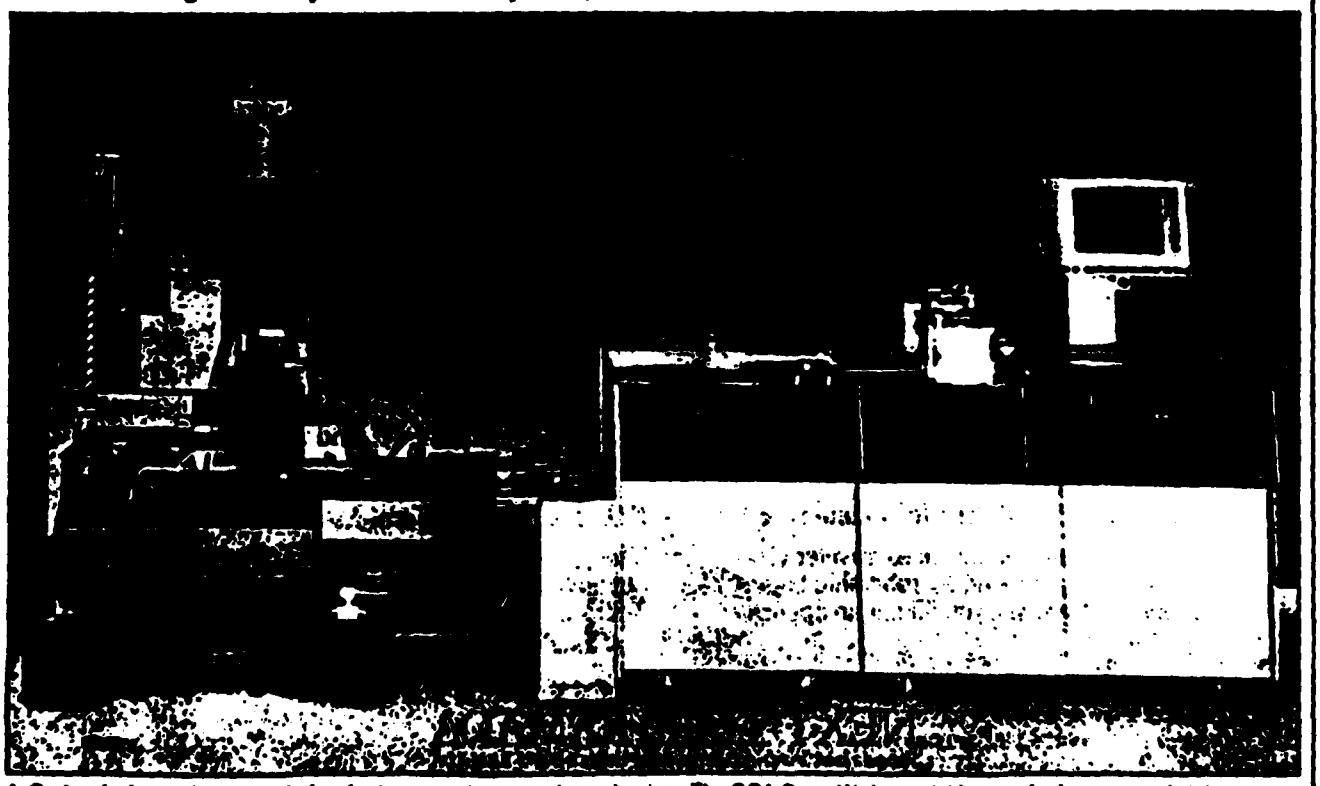
This is an evolutionary step that will automate the basic tools of IC processing and wafer-handling as well as tie together all aspects of IC production in a computer-integrated-manufacturing network.

By now the basic tools of IC processing—the lithographic, etching, and sputtering equipment, diffusion furnaces, and ion implanters—are almost fully automated. Optical lithography in particular is now heavily automated. Many commercially available systems, such as the Wafertrack (Fig. 1) from GCA Corp., Bedford, Mass., link all the lithography process's major steps—cleaning, coating, baking, alignment, exposure, development, and curing—in compact systems that move wafers on a track connecting all the major modules of the yellow (lithography) room. Each module can have an individual controller that talks to a host controller.

In the critical field of dry (plasma) etching, the latest machines, whether for single-wafer or batch processing, are completely automated with a built-in microcontroller and an interface to a host computer. Single-wafer machines such as the Omni-Etch 20,000 from Perkin-Elmer Corp., Norwalk, Conn., are fed wafers from a track transport or in a cassette-to-cassette operation in which unprocessed wafers are fed into the etcher and processed wafers exit in a cassette.

Automating a batch-etching system requires the use of either a robot or a built-in mechanism to transfer wafers from a cassette into special trays mounted in a six-sided processing chamber. The model 8300, the latest batch etcher from Applied Materials Inc., Santa Clara, Calif., is totally automated; all its operations take place in a vacuum. This makes for a high-throughput, low-particulate machine.

Inside the 8300's vacuum chamber, a three-axis robot removes wafers from a standard cassette and places them in cavities in the system's hexode chamber. At the end of an etching cycle, the robot removes the wafers and places them back in the cassette.



1. On track. A popular method of wafer transport is the track mechanism. The GCA Corp. Wafertrack lifts a wafer by means of air jets.

The latest generation of sputtering machines is almost at the same level of technology as the dry etcher. The problems are also similar: automatic handling of wafers and cassettes, process monitoring and control, and electrical and software interface to the local and host computers.

Automated wafer loading into sputtering systems from cassettes takes several forms. Hard automated systems for silicon wafers have been on the market for a number of years. Multiple cassettes on a single system reduce the number of times an operator must return to the system for loading. A minimal-complexity system is the model 202 (Fig. 2) from Materials Research Corp., Orangeburg, N.Y., in which wafers are returned to the same cassette after sputtering.

A more recent sputtering-system approach is the flexible automatic handling of different-sized substrates by robots to maintain a low particulate level. For example, on Materials Research's model 662 (Fig. 3), a robot handles silicon wafers ranging in size from 2 to 6 in. The robot is a PRI 1000 designed by Precision Robots Inc., Woburn, Mass., for clean-room service: it sheds 10 to 20 times less particulate matter than a human operator. A flat vacuum wand removes wafers from a cassette. The robot vertically orients the wafers and places them on a pallet in the sputtering system.

Automating the furnace

The diffusion furnace is one of the last IC-processing stages to be fully automated, a result of the complex wafer-handling process that the horizontal diffusion furnace requires. In this type of furnace, a wafer must be taken from a standard plastic cassette and transferred to a slot in a special quartz carrier or wafer boat. This carrier in turn must be inserted onto a paddle (quartz or silicon carbide) in a quartz tube and guided into the heating chamber. After processing, the whole procedure must be reversed.

Typical of the next generation of diffusion furnaces that will completely automate the loading and unloading of process quartzware is the Bruce Systems line from BTU Engineering Corp., North Billerica, Mass., which was shown at the recent Semicon West in San Mateo, Calif (Fig. 4).

In this system, the operator or a robot moves cassettes loaded with wafers through an input port. Inside, a mechanism or robot automatically transfers the wafers to the quartzware. The loaded boat is then moved into position and picked up by an elevator system that feeds the adjoining diffusion equipment.

After processing, the wafers are returned automatically to their cassettes in exactly the same positions in which they arrived. The cassettes are then held for removal at an output port.

Boats only

Another feature of the Bruce system and of practically all new furnaces is a boats-only soft-landing system. This system automatically loads and unloads the furnace tube with a vertically movable cantilever paddle and an automatic tube-closure mechanism. The paddle carries wafer boats having special soft-landing feet into the tube, lowers them gently until the tube is supporting them, and then withdraws. When processing is complete, the cycle is reversed.

This transport system reduces particulate generation by eliminating the sliding or rolling characteristic of wheeled paddles during insertion and withdrawal. In addition, it extends paddle life, speeds diffusion-equipment response, and saves energy by reducing the thermal mass of the paddleless cassette.

An interesting variation on diffusion furnaces is the vertical type made by only two companies, Disco-Sier USA Inc., Plano, Texas, and Tempress, a unit of General Signal Corp., Santa Clara. Disco-Sier offers an existing unit; Tempress has a brand new one.

In this type of furnace, a robotic arm transfers wafers from a standard cassette to a vertical quartz carrier on an elevator that moves the cassette vertically into the furnace. In the Disco-Sier unit wafers travel downward; in the Tempress unit they travel up into the system's furnace chamber.

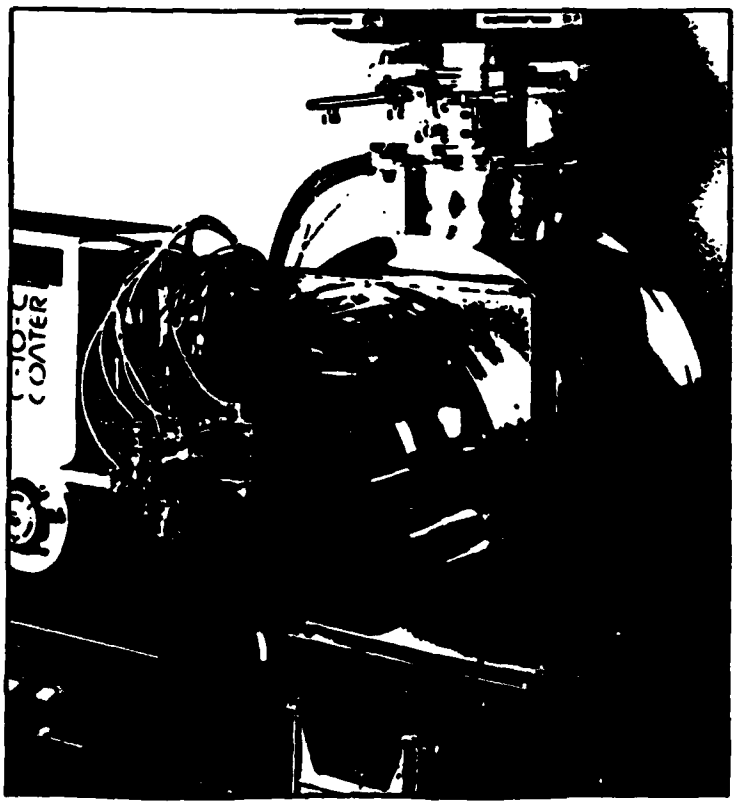
The vertical unit eliminates the quartz-to-quartz contact of the horizontal types as well as the need for soft landing systems. Its simpler robotics and horizontal transfer of wafers reduce clean-room space. Horizontal furnaces already dominate wafer-fab facilities in the U.S., but vertical furnaces are starting to show up as well.

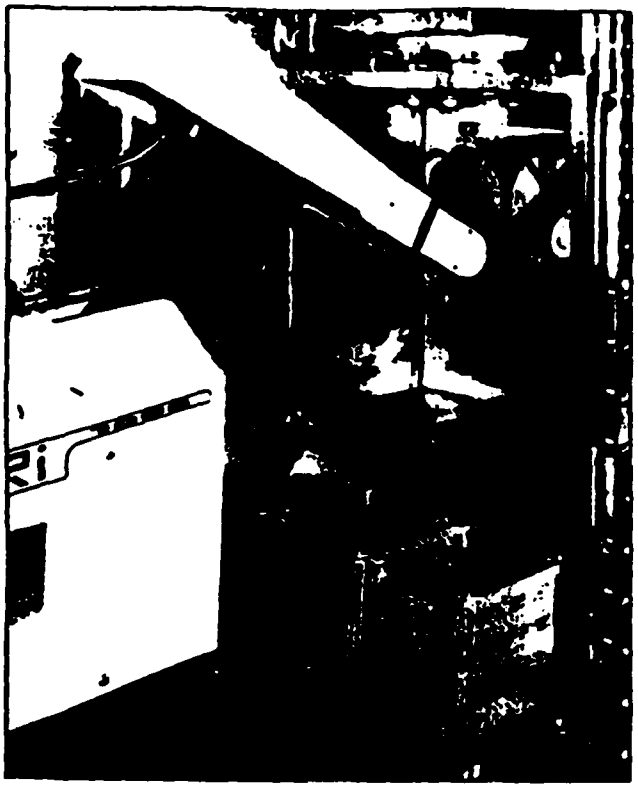
"Until recently, the tweezer or hand-held vacuum wand was the preferred wafer-handling technique for servicing most front-end process and inspection machines," notes Brian Hardegan, product engineer at Brooks Automation Inc., North Billerica. Because wafer sizes are increasing as geometries decrease, these industry practices have had to change dramatically in recent years.

Silicon-wafer manufacturing and front-end clean-room standards have generally increased by an order of magnitude, and human operators have been identified as a major source of contamination. Larger wafers are heavier, more valuable, and therefore more difficult to transport. As a result of all these factors, Hardegan and most industry experts conclude that automated wafer transport is an absolute requirement on the IC-processing line.

Several years ago, observers thought that wafers would be moved from process step to process step by track mechanisms such as belts, air jets, or walking beams. Although IBM Corp. has built such a line—known as Quick Turn Around Time or QTAT—at its East Fishkill, N.Y., facility, this technology has never

2. Wafer handling. Materials Research Corp.'s model 202 sputterer transfers single wafers from a cassette (bottom center) to a chamber (center) where wafers are processed singly.





3. Robot fed. Robots provide flexibility in handling wafers of varying sizes. A Precision Robots Inc. model 1000 robot puts 6-in. wafers into the chamber of a Materials Research sputtering system.

completely caught on. It is used most extensively for linking process modules in the lithography process.

What seems to be evolving instead is a form of automated cassette transfer that feeds single wafers serially to islands of automation. This can be done in any of four ways.

Clean-room workers generally do the direct transfer; in some cases, a multiaxis robot has sufficient placement accuracy and lifting capability. The cassette-bus concept offers increased system flexibility for routing to various process stations by transporting individual cassettes within a "clean tunnel." In-process cassettes of wafers are placed onto a vehicle that delivers them to the appropriate station, where a bar code reader identifies them and the resident robot transfers them to the next process.

The cassette shuttle is flexible. As with the bus system, a dedicated robot at each process station must transfer the cassettes to and from the tunnel.

Using sealed containers, clean-room personnel can carry the in-process cassette between work stations without significant contamination of wafers en route.

In the direct transfer of cassettes, GCA has demonstrated an extended-reach gantry robot that transfers a wafer cassette from a GCA stepper to a GCA triode etcher within a clean room.

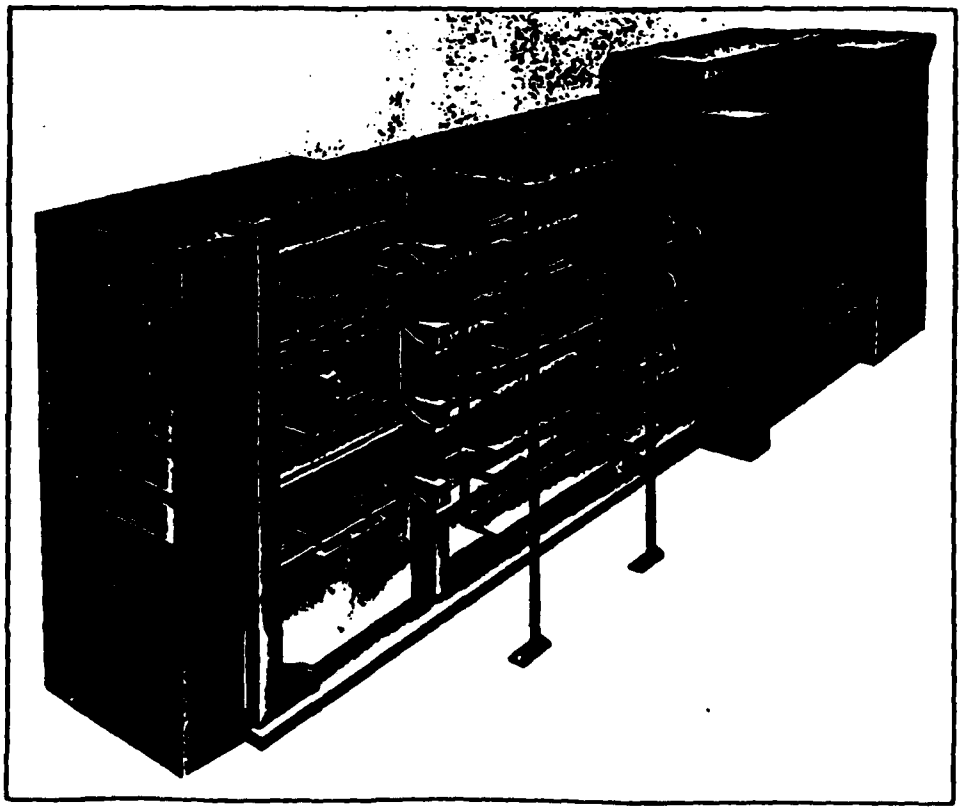
The bus concept, combining automatically guided vehicles with a built-on robotic arm for transferring cassettes, may be most likely to succeed. Currently, Veeco Integrated Automation, Dallas, and Flexible Manufacturing Systems Inc., Los Gatos, Calif., make this type of equipment. According to James Harper, president of Veeco Automation, the Japanese are heavy users of guided vehicles in wafer fabrication but without built-on robotic arms.

Guided vehicles

Veeco is the U.S. leader in this field with three types of vehicles: the V1, with no robotic arm; the V2, with a four-axis arm; and the V3, with a seven-axis arm for loading and unloading process machines. The Veeco vehicles, designed to work in a clean room, navigate by following a fluorescent paint or tape pattern on the factory floor and are commanded by an infrared communications link. The firm has already sold systems to Fairchild Camera & Instrument Corp., Mostek, and the Siemens/Philips joint VLSI venture.

Flexible Manufacturing's transport vehicle (Fig. 5) has a six-axis Intelledex robotic arm and is controlled by an IR communication link. It uses sonar for collision avoidance and does not have to be constrained by the taped pattern; it can be moved in a more flexible manner by means of a gyro-based inertial navigator. Flexible Manufacturing's goal is to achieve a 1,000-h mean time between failures with this vehicle. This company's first delivery will be to a large telecommunications firm in Florida.

Another strong contender in cassette-transfer technology is a technique developed at Hewlett-Packard Co., Palo Alto., just last year—the Standard Mechanical Interface, or SMIF, now being made at Asyst Technologies Inc., Fremont, Calif. SMIF systems (Fig. 6) involve small, sealed, dustproof boxes for



4. Happy landings. A new Bruce Systems diffusion furnace delivers wafers from cassettes to quartzware and then returns them to their identical cassette positions after processing is complete.

storing and transporting wafer cassettes. Each box has a specially designed "door" and each enclosure or canopy surrounding the equipment has a mating door. This mating arrangement is the Standard Mechanical Interface. The environment inside this enclosure or canopy is, in effect, a miniature clean room.

In transferring cassettes, the two doors open simultaneously, trapping particles that may have been on the outside surface of either door in the space between them. The equipment must be fitted with a special mechanism that operates the doors and transfers the cassette between the box and the equipment's indexer. This transfer has to be performed so that the external environment does not contaminate the wafers.

The SMIF system is flexible because either human or robotic-operators can transport its boxes. HP has demonstrated particle reduction by as much as 10 times that possible in a conventional clean room.

A recent variation of the SMIF system involves SMIF boxes and a special-purpose robotic mechanism interfaced to a stepper. SMIF mechanical interfaces have already been designed into semiconductor-processing equipment from MTI, Tylan, and GCA.

If all the equipment in a yellow room is put under individual enclosures, each with a locally controlled environment of Class 10 or better, it is possible to run the room overall at a Class 1000 level. This full SMIF implementation can significantly reduce space and energy consumption. For example, it would be possible to reduce the 250,000 ft^3 of space in a typical clean room to 2,500 ft^3, with the concomitant cost and energy savings.

Automation by itself is insufficient to get to the overall objectives of high yield, low costs, and increased productivity. A modern semiconductor facility needs a computer-integrated-manufacturing network to manage, monitor, control, collect, and interpret data, and to link all elements of the fabrication process with a communications network. Thus, a whole new CIM software and hardware industry has sprung up just to support IC production.

5. On the move. Guided vehicles such as the FMS mobile transport vehicle, will shortly be carrying cassettes of wafers between processing tools on automated IC fabrication lines.

The main players in this new industry are I.P. Sharp Associates, Toronto, Ont., Canada; HP; Sentry/Schlumberger Computer Integrated Manufacturing Systems, Los Altos, Calif.; Consilium, Mountain View, Calif.; and CTX International, Sunnyvale, Calif. Some newcomers are GCA, Isitec Corp., BTU Engineering, and Kulicke and Soffa, which has a CIM system for the assembly end of an IC line.

I.P. Sharp's CIM software is called Promis and was developed under a joint effort by the Canadian company and General Electric Co.'s Solid State Operation. It is currently installed at 16 companies worldwide. Promis uses a distributed-systems architecture, with microprocessors performing the front-end facilities-monitoring functions and a Digital Equipment Corp. VAX computer to coordinate control of the entire manufacturing process. The Semiconductor Equipment Manufacturers Institute's Semiconductor Equipment Communication Standard 1 and 2 communication protocol links the processing equipment to the host computer.

Promis cuts paperwork

Silicon Systems Inc., Tustin, Calif., a custom and semicustom IC producer, has been running Promis for three years with a dual VAX setup and more than 40 terminals. Currently, the CIM software is eliminating paperwork and is controlling diffusion furnaces and automatic test equipment as well as specifying recipes. Future plans are to interface the facilities-monitoring software to lithography-inspection equipment. "We feel process problems picked up by this system have paid for the cost of Promis," emphasizes Larry Cleland, manager of manufacturing systems planning at Silicon Systems.

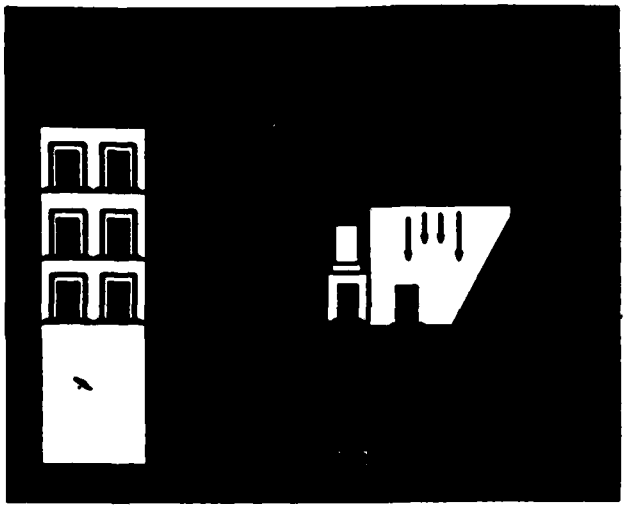
Hewlett-Packard's CIM system is called the Semiconductor Productivity Network (SPN), developed as a distributed-computer network installed to control one of the firm's IC facilities (*Electronics*, June 5, 1980, p. 151). The SPN is based on two HP computers, the models 1000 and 3000. The 1000 supervises real-time equipment functions while the 3000 performs transaction-processing functions and data-base management. Usually, multiple HP 1000s are networked with a central HP 3000. HP has 10 software modules, including a TC-10 tester collection system, a PC-10 process-control system, and PL-10 production-planning system.

Hewlett-Packard has 34 customers using SPN at more than 90 sites. One of the largest is Intel Corp. The Santa Clara firm has realized significant improvements in product yield with computerized data collection, tracking, analysis, and control systems designed for the semiconductor industry. Intel is installing SPN software to run on HP 3000 computers located in its eight semiconductor-wafer-fabrication facilities and five assembly-and-test facilities around the world.

The SPN system has already made a difference at Intel's beta-test site. Early in 1984, a process engineer made an adjustment on a mature product—a mask critical-dimension target change—and managed to increase the product sort yield by 10%. Robert B. Clifford, Intel's manager of components data automation, estimates that this slight adjustment paid for the plant's entire system, both hardware and software, in just four months.

Sentry/Schlumberger's entry into the CIM arena is called Incyte II. This is a complete CIM package for fabrication, assembly, and test built around a VAX. It has its own local-area network based on Ethernet and completely supports the SECS protocol.

The CIM network can cover the entire semiconductor production flow from silicon-ingot growth to assembly and final test. Incyte is installed at 12 Fairchild Semiconductor sites in North America and Europe and has led to 5% to 7% improve-



6. SMIF. The Standard Mechanical Interface concept isolates the wafer during storage and transport. Particle-free SMIF boxes are transported to special SMIF transfer mechanisms that send the cassette into equipment with special enclosures.

ments in yields; cycle times dropped by 23%, and products needing rework declined by as much as 30%.

Consilium's software, called Comprehensive Online Manufacturing and Engineering Tracking System (Comets), is also based on a VAX and uses the SECS protocol. The Palo Alto firm's software package includes 12 modules for such functions as work-in-progress tracking, engineering data collection, nonlot tracking, data analysis, factory communications (for a paperless environment), process-automation management, facilities monitoring, capacity planning, and scheduling.

For the past year and a half, Harris Semiconductor Corp., Melbourne, Fla., has used two Comets software systems to control a diffusion furnace and a lithography inspection machine. The CIM software has reduced lot-traveler records (paperware that moves with each discrete lot) from 20 pages to 1 page—an almost paperless operation. It is still too early to gauge the system's effect on yield and costs, but Harris's CIM experts find that the software's wafer-tracking and record-keeping features have minimized scheduling problems.

With the exception of Incyte, all the above IC CIM systems do not deal with the assembly end of IC production, which consists of wafer preparation and dicing, die bonding, wire bonding, encapsulation, and testing. At Semicon West, Kulicke and Soffa Industries Inc., Horsham, Pa., introduced a factory-automation system called the Semiconductor Assembly Management System to fill this void. Its goal is to increase the productivity and quality levels of semiconductor assembly.

Three-tier architecture

Kulicke and Soffa's network has a three-tier architecture that distributes intelligence from the machine level through data concentrators and up to a central factory host computer. The lowest tier of the network is the peripheral communications controller, a single-board computer designed around a Motorola 6809 processor. Each assembly machine in the automation system needs its own peripheral controller. Operating under SEMI's SECS protocol, these boards transmit data to the higher-level system through RS-232-C ports. Error codes, bond tables, wafer maps, machine processing data, material identification, and tracking information can be communicated through this system.

At the system's middle level, an intelligent data concentrator consolidates the information received from a group of machines. The concentrator is a DEC MicroVAX or equivalent, which can support up to six pieces of assembly equipment. The concentrator then communicates information to the system's uppermost tier, which is the factory floor controller, over Ethernet.

The controller can range in size from a DEC VAX 11/730 to a VAX cluster. The controller in turn can interface with a higher-level management information system. This host would provide overall scheduling information, data from prior assembly operations or wafer fabrication, and the basic coordination for the overall assembly operation.

The real thing

Although all the basic tools of the technology are now available, with the exception of IBM's QTAT line full automation of IC-production lines in the U.S. is just beginning. But it could halt for a while: the current recession has caused cancellations and delays of many planned automated facilities. Japan is much farther along in full IC-processing automation.

In their Semicon West paper on the impact of process automation on VLSI-chip yields, Masakatsu Nakamura, Kanro Sato, and Yasuske Sumitomo of Toshiba Corp. define three stages of IC automation.

The first is automatic operation of main equipment in the cassette-to-cassette mode with an automatic operation sequence using microcomputer control. The second is local in-line systems for main processes, automatic wafer handling and transportation, and local computer networks for process and lot control.

The third stage of IC automation connects processing areas by means of vehicles and automates the transfer from one piece of equipment to another by robotics. It also automates visual checks of geometry, cross sectional structure, particle density and other physical parameters, as well as links complete computer networks for process and production control (CIM).

Currently, most U.S. firms are still in stage one, with many already having the computer network of stage three. In general, stage three is only in the planning stage.

In Japan, many companies already are at stage three. Toshiba is one of the few that will talk about its system in detail.

Toshiba has developed a new computer-aided-manufacturing technique that approaches stage three. The technique's main purposes are production control and quality control. The line includes a computer network for production and process control, in-line systems for main processes, automatic wafer handling, automatic transportation, and monitoring systems for particle detection.

Toshiba's CIM network uses a three-layer hierarchy. The top layer consists of a host computer for managing the whole system. Next comes a block computer, which manages communication between the host and the bottom layer of machine controllers. Communication between block and machine controllers is standardized using the SECS protocol.

Toshiba uses an automatic transportation system that carries a load of wafers from one process room to another. This system is controlled by the block computer, which gets its destination commands from the host. Transportation between rooms is by a monorail vehicle that hangs from the ceiling and passes above all other equipment. (There are elevators at each station.) Switches in the overhead rail direct vehicles to different destinations.

Automation brings its own problems along with its benefits. Toshiba's Nakamura, senior manager in the Integrated Circuit Advanced Process Engineering Department, IC Division, notes that mechanical automation can replace some handling operators but not the maintenance engineers who bring in a lot of contaminating particles. Therefore, equipment with high reliability and short maintenance time is essential for an automated plant. Toshiba's solution is simple structured equipment with primarily modular parts. □

END

FILMED

MARCH, 1988

DTIC